Proceedings of the 5th International Conference on Meaning-Text Theory

Barcelona, September 8 – 9, 2011

Igor Boguslavsky and Leo Wanner (eds.)

ISBN 978-84-615-1716-9

CONTENTS

Foreword

We are pleased to present the proceedings of the 5th International Conference on Meaning-Text Theory (MTT 2011), to take place on September 8th and 9th 2011 at the Pompeu Fabra University, Barcelona. Compared to the previous MTT Conferences, MTT 2011 introduces two novelties. For the first time, an MTT Conference is held in conjunction with another major event, namely the International Conference on Dependency Linguistics (DepLing). This was conceived to make the MTT Conference more visible and attractive to researchers who work on related theories or related contexts, but were not aware so far of this proximity. Looking at the number of submissions from people outside the inner MTT circle, we are inclined to conclude that we succeeded. However, the time will show whether we can maintain this opening up in the future. Also for the first time, the Program of an MTT Conference is organized in terms of six thematic tracks: Fundamentals, Dictionaries and Lexical Semantics, Collocations, Semantic Derivation and Morphology, Computational Applications, and Terminology. This was to better accommodate the conference to the growing diversity of the themes addressed by the MTT community and to further open the conference to researchers not acquainted with MTT as a whole, but interested in a specific thematic area covered in MTT. The outcome of this experience is somewhat mixed. While some of the tracks (e.g., Fundamentals and Dictionaries and Lexical Semantics) attracted a large number of submissions, others (especially Semantic Derivation and Morphology and Terminology) received a disappointingly low number of submissions. However, we are confident that these are merely startup challenges that will be overcome in the future editions of the conference. Overall, we can state that the total number of submissions exceeded our expectations; the acceptance rate has been this time somewhat lower than 70%.

As the proceedings of the previous MTT conferences, the MTT 2011 proceedings are being published electronically and can be accessed at http://www.meaningtext.net.

We would like to thank our invited speakers Ignacio Bosque and Leonid Iomdin and all the participants for making MTT 2011 a stimulating scientific forum. Furthermore, we would like to extend our thanks to the Track Chairs Amparo Alcina, Margarita Alonso Ramos, Valentina Apresjan, Maarten Janssen, Sylvain Kahane, Marie-Claude L'Homme, and Tilmann Reuther and to the Program Committee for their efficient and thorough work on paper selection. Many thanks also to Joana Clotet and Bea Abad for taking care of practically all matters related to the local organization of the conference, to Simon Mille for assisting them and to all the other members of the local organization team: Stefan Bott, Alicia Burga, Gerard Casamayor, Gaby Ferraro, Estela Mosquiera, Luz Rello, and Orsi Vincze. Financial support for MTT 2011 was provided by the Natural Language Processing research group TALN of the Pompeu Fabra University (UPF), the Department of Communication and Information Technologies, UPF, the Department of French and Romance Philology at the Autonomous University of Barcelona, and Inbenta.

Igor Boguslavsky and Leo Wanner

PROGRAM COMMITTEE CHAIRS

Igor Boguslavsky, Russian Academy of Sciences and Polytechnical University of Madrid
Leo Wanner, ICREA and Pompeu Fabra University, Barcelona

TRACK CHAIRS

Amparo Alcina, Jaume I University, Castellón
Margarita Alonso Ramos, University of La Coruña
Valentina Apresjan, Russian Language Institute, Russian Academy of Sciences
Igor Boguslavsky, Russian Academy of Sciences and Polytechnical University of Madrid
Maarten Janssen, Institute for Applied Linguistics, Pompeu Fabra University, Barcelona
Sylvain Kahane, Modyco, University Paris Ouest
Marie-Claude L'Homme, OLST, University of Montreal
Tilmann Reuther, Alpen-Adria University, Klagenfurt
Leo Wanner, ICREA and Pompeu Fabra University, Barcelona

PROGRAM COMMITTEE

Lorraine Baqué, Autonomous University of Barcelona
Maria Auxiliadora Barrios Rodriguez, Complutense University of Madrid
David Beck, University of Alberta
Bernd Bohnet, University of Stuttgart
Margarita Correia, ILTEC, Lisbon
Béatrice Daille, LINA, University of Nantes
Dmitry Dobrovolsky, Russian Language Institute, Russian Academy of Sciences
Pamela Faber, University of Granada
Kim Gerdes, Sorbonne Nouvelle
Boris Iomdin, Russian Language Institute, Russian Academy of Sciences
Lydia Iordanskaja, University of Montreal
Anne-Laure Jousse, Druide Informatique Inc.
François Lareau, Macquarie University
Roman Laskowski, Polish Academy of Sciences
Patrick Leroyer, Aarhus Business School
Irina Levontina, Russian Language Institute, Russian Academy of Sciences
François Maniez, University Lumière Lyon 2
Igor Mel'cuk, University of Montreal
Jasmina Milicevic, Dalhousie University, Halifax
Alain Polguère, University Nancy 2
Agnès Tutin, Université Grenoble 3
Serge Verlinde, Catholic University of Leuven
Daniel Weiss, University of Zurich
Robert Zangenfeind, University of Munich

LOCAL ORGANIZATION OFFICE

Joana Clotet, Pompeu Fabra University
Beatriz Abad, Pompeu Fabra University
Simon Mille, Pompeu Fabra University

LOCAL ORGANIZATION CHAIR

Lorraine Baqué, Autonomous University of Barcelona

LOCAL ORGANIZATION COMMITTEE

Stefan Bott, Pompeu Fabra University
Alicia Burga, Pompeu Fabra University
Gerard Casamayor, Pompeu Fabra University
Angels Catena, Autonomous University of Barcelona
Gabriela Ferraro, Pompeu Fabra University
Estela Mosquiera, University of La Coruña
Luz Rello, Pompeu Fabra University
Judith Sastre, Inbenta Inc.
Orsolyia Vincze, Pompeu Fabra University

INVITED SPEAKERS

Ignacio Bosque, Complutense University of Madrid and Spanish Royal Academy
Leonin Iomdin, Russian Academy of Sciences

Deducing collocations

Ignacio Bosque
Departamento de Lengua Española, Facultad de Filología-D
Universidad Complutense, 28040 Madrid
ibosque@filol.ucm.es

Abstract

There is a certain consensus on some defining features of collocations (they are frequent, recurrent, conventional, institutionalized, etc.), but there is not much agreement on other characteristics, for example whether or not they constitute binary lexical relations, and, crucially, whether they are arbitrary, and must be stipulated, or should instead be deduced or inferred. I will argue that most collocations are not instances of binary relations, and also that these combinatorial associations are not memorized as individual pieces of lexical information. Collocates should not be stipulated or specified individually, since the appropriate bases for them constitute large paradigms which meet a number of restrictive semantic criteria. Speakers (whether native or not) who learn collocations have intuitive access to these abstract semantic features, which are proven to be recurrent throughout the grammar. Lexical functions in Meaning-Text Theory (MTT) could also be formulated taking these semantic groups as bases, instead of the specific keywords provided by their respective paradigms.

Keywords

Collocations, explanatory and combinatorial lexicology, lexical functions, semantic compatibility.

1 Features of collocations

One of the first things that one learns about the field of collocations is the fact that there is a set of basic, defining features of these peculiar word combinations that is often repeated in the abundant literature of this topic. It seems to me that these features come in two groups: non-controversial features and controversial features. Non-controversial features of collocations do not seem to be very interesting, since they do not provide crucial information about them, but controversial features are quite useful, I believe, even if some of them are problematic, or simply wrong. The features that I consider to be non-controversial are the following:

NON-CONTROVERSIAL FEATURES OF COLLOCATIONS:

1. Recurrent
2. Frequent
3. Conventional
4. Institutionalized.

The fact that collocations are recurrent is a natural consequence of their being systematic. Since grammatical and lexical phenomena are expected to be recurrent, it seems to me that there is nothing special about 1. As for 2, frequency is, again, a consequence of systematicity. Even if one takes collocations to be more frequent than other expected combinations, we should evaluate expectation in order to understand this property.

We must also be aware of the fact that reasons for frequency are often extra-linguistic. Let me give a simple example of this. According to the *Collins Cobuild English Collocations on CD ROM* dictionary (CCEC), based on the *Bank of English Corpus*, the English nouns most frequently combined with the verb *protect* are *people, interests, order, rights* and *children.* I do not have any reason to question this statement, but notice that the relevant question is whether or not we should take it to be proper "linguistic information". Certainly, frequency of word combinations (whether linguistically grounded or not) is a useful tool for automatic disambiguation and other tasks in natural language processing, but notice that those are technical applications of the basic data that we try to understand. The frequencies that the *Cobuild* dictionary provides for nouns combined with the verb *protect* hardly qualify as linguistic information, in a strict sense, for the fairly obvious reason that, from a linguistic point of view, you can protect anything. In fact, this information is more interesting for a sociologist than for a linguist: it is information about our customs or our habits. It tells us something about the frequency of some human actions, rather than the combinatorial properties of words.

We may admit 3 as well, but granting that conventional phenomena are part of many morphological or syntactic patterns, thus not a particular feature of collocations. As regards 4, another alleged feature of collocations, notice that the adjective *institutional* does not apply to linguistic phenomena (whether morphological, syntactic or lexical) in a straightforward manner. I do not mean that it is wrong, but rather that it is scarcely informative as a distinctive characteristic of collocations, as opposed to other linguistic patterns.

In this talk I would like to concentrate on the controversial features of collocations, which I take to be the following:

CONTROVERSIAL FEATURES OF COLLOCATIONS:

1. Salient
2. Binary
3. Arbitrary
4. Transparent.

We may admit that collocations are salient combinations, but saliency can be understood as a psychological concept or a statistical notion. In the first case, it must be evaluated in relation to 3 (in the same list), since saliency will then depend on the semantic classes in relation to which restricted combinations make sense. If saliency is interpreted as a mere statistical notion, it will take us back to frequency.

I believe it is important to stress the idea that collocations should not give us information about things people do often, but about the restrictive properties of lexical combinations within a linguistic system. The combinations with the verb *protect* that I just quoted from the *Cobuild* dictionary are, without any doubt, frequent, conventional, salient and even institutionalized, but —from a restrictive point of view— I do not think they are collocations. Since

the concept of 'collocation' is nowadays used for different purposes by different people in a number of different interpretations (often incompatible with each other), one should not avoid the controversial question whether or not attested, frequent combinations based on extra-linguistic information should be properly considered to be collocations. If I am not mistaken, most Meaning-Text Theory (MTT) practitioners would answer NO to this question. Although it would be fair to say that my work in the field of collocations and related matters in the grammar of Spanish (Bosque 2001a,b; 2004a,b) does not quite comply with some MTT principles, I completely agree with mainstream MTT on this issue.

As for controversial feature 2, my belief is that this feature is wrong if *binary* is interpreted as "associated with a lexical item in a one-to-one exclusive relationship". The reason is the fact that collocates may be compatible with long paradigms of bases, an important point to which I will return shortly. Binarity would certainly be a feature of collocations in the trivial sense that a particular base is always related to some particular collocate in some particular text.

The feature *arbitrary* is also controversial, but it largely depends on the precise meaning of this term. Collocations are arbitrary if this adjective is interpreted in a Saussurean sense, but they are not if *arbitrary* is interpreted as "non-predictable", since most collocations are not memorized individually, as I will argue. As for *transparency*, the final controversial feature, it is a matter of degree (as it is often acknowledged) whether or not this is a property of collocations. In the following pages I will try to argue for a positive answer to this question.

2 Collocates and paradigms of bases

I will assume the by now widely accepted distinction between *bases* and *collocates* (Hausmann 1984, 1989). As it is well-known, collocates are analyzed in MTT as values for keywords or bases of lexical functions (LFs) (Mel'čuk, 1996). Taking this for granted, I would like to emphasize some properties of collocates seldom stressed, and often hidden (or even rejected) in the vast literature on collocations:

- Most collocates are fully meaningful. The restrictive relation which holds between bases and collocates depends, to a large extent, on that property.
- Most collocates are related to bases which constitute semantic classes. These classes are recurrent throughout the grammar. If we analyze lexical functions as specific binary relations, we will miss a large number of generalizations.
- The features that allow us to build these classes are part of the native speaker's knowledge of the language.

Here is an example. The piece of information encoded in a LF such as Magn(*increase*) = *substantially* belongs to a broader generalization, since the adverb *substantially* typically modifies verbs of change of state, more specifically so-called *gradual completion verbs* or *degree achievement predicates* (Declerck, 1979; Tenny, 1994; Bertinetto and Squartini, 1995; Levin and Rappaport, 1995; Kennedy and Levin, 2007, among many other studies). Here are the major lexical groups that we may distinguish:

1. Verb of increasing: *enhance, enlarge, exceed, expand, go up, grow, improve, increase, overshoot, progress, raise, upgrade*, etc.

2. Verbs of decreasing: *cut back, decline, degrade, diminish, discount, dwindle, fade, fall, lessen, limit, minimize, moderate, reduce, shrink, slow, tighten, weaken,* etc.
3. Verbs denoting the notion of change: *alter, amend, change, deviate, differ, diverge, fluctuate, modify, redefine, reformulate, revise, rewrite, transform, vary,* etc.
4. Comparative adverbs (*more, less*) and adjectives (*higher, lower, greater, older, bigger,* etc.).

To these groups we may add a few adjectives denoting (un)equality, such as *different, similar* or *identical,* and some other minor lexical classes to which I will return below. Notice that the verb *rewrite* appears in group 3, whereas the verb *write* does not belong to any of these paradigms. In fact, speaking of a chapter of some book, one could naturally say that it has to be "substantially rewritten", but one would not say that it has to be "substantially written". In the same vain, notice that two analyses can be "substantially identical", but not "substantially interesting", and, also, that some amount of money might be "substantially reduced or increased", but not "substantially calculated or donated". Since any speaker of English has an intuitive knowledge of these contrasts, and many other similar to them, it seems to me that the natural move would be to formulate LFs in a way that they involve the lexical classes sketched, or some variants of them, rather than specific lexical items. That is, instead of

Magn(*increase*) = *substantially*,

we would have something like

Magn(VERBS OF INCREASING) = *substantially*.

If, contrary to what it is commonly assumed, we formulate LFs relative to lexical groups or semantic features, rather than particular lexical items, we will not have to specify collocates as individually choices of hundreds, perhaps thousands, of bases. It is in this particular sense that a large number of collocations are DEDUCIBLE. I am using the verb *deduce* here as opposed to *postulate* or *stipulate*. I simply want to stress the fact that, if we stipulate collocations as binary, individual choices (avoiding semantic groups such as the ones above), they would represent a huge amount of specific, redundant and hardly learnable lexical information.

But we must admit that semantic generalizations are never easy, sometimes not even possible. In fact, it would we appropriate to distinguish between *deducible* and *non-deducible* collocations. The difference is somehow similar to Cowie's (1983) distinction between *open* and *restrictive* collocations, but there is a terminological aspect of the latter distinction that seems unclear to me. Notice that, by opposing *open* collocations to *restrictive* collocations, we infer that the former are not restricted. But this is not correct: so-called *open collocations* are restricted as well, even if restrictions come from more abstract lexical or semantic features.

In *deducible collocations,* such as the example I just gave with the English adverb *substantially,* collocates are not individual choices for bases. On the contrary, bases form paradigms defined on semantic grounds, to which collocates are sensitive. This is, in fact, a standard characteristic of most predicate-argument relations.

We must also acknowledge the existence of *non-deducible collocations*. These form reduced (*i.e.* closed) paradigms of bases for each collocate, or rather no paradigm at all. They include

examples such as the English verb *bleat* (with *sheep* or *goat* as possible subjects), or perhaps VPs such as *pay someone's* Ns *to someone else*, where Ns stands for a reduced paradigm of nouns containing *respects*, *regards* and perhaps a few other. The Spanish VP *conciliar el sueño* ('get to sleep') is also a good example of this pattern, since this is not a VP idiom (as opposed to, say, *tomar el pelo* 'pull one's leg') but, even so, it does not give rise to a paradigm. In the case of the verb *fruncir* ('frown'), a very short one is obtained: *ceño, entrecejo, cejas* ('brows' the three of them, with minor differences).

Sometimes, paradigms of bases for a single collocate are rather short, but not necessarily closed. Note that an LF such as LiquFunc_0(*order*) = *countermand* is straightforward, but it somehow misses the point that there is a paradigm of bases here (*order, command, request, instruction, sanction*, etc.). The existence of the paradigm can be simply thought as the natural consequence of the fact that *countermand* has some meaning that we should grasp. The paradigm is not necessarily closed, since a person might countermand somebody else's execution or some other similar action.

Light verb constructions fall somehow in between deducible and non-deducible collocations, since they give rise to paradigms of bases with exceptions. Possible example include English *take a bath, take a shower, take a dip, take a plunge, take a swim*, etc., or Spanish *dar un paseo, dar una vuelta, dar un garbeo* (all three, 'take a walk'), *dar un rodeo* ('make a detour'), but not *dar *una excursión* ('go on an outing').

Not many scholars have emphasized the need to search for generalizations on collocates, but some of them have noticed the relevance of this way of looking at them. Outside MTT, I have in mind Aisenstadt (1979), Cowie (1979, 1981, 1983), Stubbs (1996, 2001) or Nesselhauf (2005). Cowie (1978, 1981) sometimes characterizes "collocational ranges of open collocations" on semantic grounds, as in *Run*: INSTITUTION, ORGANIZATION (*run a business, a theatre, a bus company*). In other occasions, he simply provides the relevant paradigm with no covering semantic label, as in Cowie (1981: 227):

> Entertain: *idea, notion, suggestion, proposal, doubt, suspicion, …*
> By chance: *find, discover, notice, come across, meet, happen, come about, …*

Inside MTT, I think that J. Apresjan is perhaps the scholar who has made that point in a most straightforward manner (Apresjan, 2009; Apresjan & Glovinskaja, 2007). For example, Apresjan & Glovinskaja (2007) present a series or proposals on the semantic grouping of bases for LFs. Their examples include the following:

- *Commit* is preferred as the value of Oper_1 from nouns denoting negatively evaluated acts: *commit aggression, a blunder, a crime, an error, murder, a sin, suicide, treachery.*
- *Draw* is preferred as the value of Oper_1 from nouns denoting a mental operation, especially various operations of comparison: *draw an analogy, a comparison, a distinction, a parallel* (vs. *draw a conclusion*).
- *Enjoy* is preferred as the value of Oper_1 from nouns denoting those social attributes of people which can be seen as privileges: *enjoy freedom, privileges, a good reputation, rights* (vs. *enjoy good health*).

> *Experience,* and especially *feel,* often collocate with nouns denoting emotional states and attitudes: *experience grief, joy, pleasure* (vs. *experience need*); *feel fear, hatred, pity*.

Moreover, they show that different groups are possible for the same collocate. Here is one of their examples:

> *Undergo* as a value of $Oper_2$ is preferred with three groups of nouns:
> (a) Nouns denoting hostile actions or activities: *undergo criticism, an interrogation, punishment, torture*;
> (b) Nouns denoting various forms of inspection: *undergo censorship, a check-up, an examination, inspection, a test;*
> (c) Nouns denoting interference: *undergo an operation, reform, repair, surgery*, etc.

This is exactly the approach taken in the combinatory dictionary *REDES,* which I coordinated some years ago. This is a dictionary of Spanish in which the semantic groups that bases constitute are studied in detail for a large number of collocate (see also DeCesaris and Battaner, 2006, for an outline of its main lexicographic features). In the rest of this paper I will point out the major advantages of this approach to collocations, and also some problems and perspectives that this line of research gives rise to.

3 Some advantages of the deductive approach to collocations

3.1 Learning

Paradigms of bases can be easily extended, as long as they fit the appropriate features. The speaker is not supposed to memorize long lists of verbs denoting 'increasing', 'decreasing', 'change', etc, in the case of the adverb *substantially*, or other lists of bases for similar collocations. These combinations are not memorized individually by native speakers, regardless of the obvious fact that computers may have quick access to those extensive paradigms in a trivial way.

3.2 The relevance of the collocate's meaning

Notice that the definition of the verb *read* seems to play no direct role in either the formulation of a LF such as $Real_1$(*book*) = *read* or the linguistic characterization of the possible nominal bases for $Real_1$(Noun) = *read*. But the meaning of collocates (either as defined by dictionaries or in a different format) is an important piece of evidence in the analysis of collocations.

If one accepts the reasonable assumption that a LF such as $Real_1$(*prediction*) = *corroborate* is not a specific piece of lexical information that the learner has to memorize, two natural questions will arise: (i) whether or not the possible nominal bases in the LF $Real_1$(Noun) = *corroborate* have something in common; and (ii) whether or not the definition of the verb *corroborate* ("to strengthen, support or confirm something with additional evidence") plays any role in the task or restricting this nominal paradigm. I believe that the answer to both ques-

tions in affirmative. Here is a tentative classification of the possible nominal bases for the verb *corroborate*:

1. NOUNS DENOTING PROPOSED EXPLANATIONS: *account, analysis, calculation, claim, diagnosis, hypothesis, interpretation, theory, version, view, …*
2. NOUNS DENOTING RESULTS GENERALLY OBTAINED FROM INQUIRY OR RESEARCH: *conclusion, confession, evidence, finding, observation, report, result, testimony, …*
3. NOUNS DENOTING RESULTS OF SUPPOSING OR EXPECTING SOME INFORMATION: *rumor, supposition, suspicion, …*
4. NOUNS DENOTING OTHER PIECES OF INFORMATION: *information, data, story, words, statement, detail, …*

These nominal bases give rise to about 50 collocations, perhaps some more. Again, it is very unlikely that speakers of English, where native or not, learn them by heart on individual basis.

The relevance of the collocate's definition in the task of restricting its possible bases is easily confirmed by simple LFs such as Magn(*seal*) = *hermetically*, where *hermetically* approximately means "in a way that air cannot escape or entry", or even Magn(*increase*) = *substantially*, since a standard definition of the adverb *substantially* is "to a great extent or degree". In fact, processes of change of state and other so-called *gradual completion verbs* nicely fit in the paradigm suggested by this simple definition (Kennedy and Levin, 2007). Many other similar examples could be added here.

We may now wonder about the precise way in which these lexical classes are related to the collocate's definition. This is a very relevant question which we could not raise in the *REDES* project. On theoretical grounds, one should expect that the lexical features involved in the semantic groups in which bases are gathered are also present in the description of the collocate's meaning. If this could be proved for a significant number of collocations, it would certainly be a most welcome result.

3.3 Deriving the information provided by dictionaries of collocations

This advantage of the deductive approach is not a minor one. A large part of the information found in dictionaries of collocations (BBI, OCD, MCD, etc.) could be automatically obtained if entries of collocates were based on semantic classes. This is exactly what we did in the combinatory dictionary *REDES*, which, as I have pointed out, was conceived as a dictionary of collocates, rather than a dictionary of bases. In fact, entries for bases in this dictionary are interpreted as simple indexes which reverse the semantic information considered to be central: the one provided in entries for collocates. Notice that the latter is precisely the kind of information that a computer program could never obtain by itself.

To build up entries for bases, we developed a rather simple system of superscripts which sends the user to the specific point of the correspondent collocate entries. For example, the entry for *victoria* 'victory' (a base) contains adjectives and verbs such as these:

abrumador[17] , *abultado*[25], *a domicilio*[21], *agridulce*[8] , etc. (many other follow in alphabetical order).

abocar(se)(a)[43], *acariciar*[2] , *aderezar*[4], *aguar(se)*[30], *airear*[42], etc. (many other follow in alphabetical order).

More generally, entries for most nouns in *REDES* (quantificational nouns are not included because they are collocates) are just indexes of the core restrictive semantic information provided in entries for collocates. This information is organized in lexical classes, roughly as in the examples above. I will return to this point below.

It is hard to avoid "the issue of direction" as applied to lexical selection; that is, the question of whether bases select for collocates or the other way around. Although I have discussed this issue somewhere else (Bosque 2004b), it may be worth pointing out here that this is just a terminological problem: the term *selection* may be used with an intentional interpretation, but also in a somehow more standard and technical non-intentional reading. In the former case, it is the speaker who deliberately chooses some lexical item to suit some particular linguistic need (for example, expressing the notion of coming into existence as applied to some abstract entity). In the latter case, a predicate selects for its arguments and, by doing so, it restricts the paradigm of its possible complements. In this interpretation, the notion of intentionality does not play any role, as standardly assumed.

Notice that, if the distinction is made, collocates turn out to be natural candidates for entries in a combinatory dictionary, since it is predicates that restrict arguments, rather than the other way around. Needless to say, the information provided in entries for bases in standard dictionaries of collocations may not be primitive, as I have tried to explain, but it is quite useful for learners and other users. The very fact that this information "follows from something else" does not have to be interesting or intriguing for common users, much less worrying. My point is simply that perhaps it should be for lexicologists.

3.4 Literal interpretations vs. figurative senses

By focusing the meaning of collocates we may show, in a very explicit form, the transition from their literal interpretation to more abstract or figurative senses. It has been repeatedly pointed out in most cognitive approaches to linguistics, but also in other frameworks,[1] that these extensions provide an important piece of linguistic information for both native and non-native speakers. If we consider simple pairs such as the following:

> *relish the tomato / relish the victory; sugar dissolves / problems dissolve; disguise a child / disguise the truth; block access / block negotiation; straighten a tie / straighten a problem,*

we may realize that the meaning of the verbs involved does not necessarily change depending on whether their nominal argument denotes some physical entity or a more abstract notion. I realize that these are long-discussed and much controversial matters in the fields of lexicography and lexicology. Even so, their relevance for the analysis of collocations is quite evident,

[1] The number of references here is too large to cite. I will simply point out a few of them (namely, Glucksberg, 2001; Fauconnier & Turner, 2002; Vega Moreno, 2007) and refer to the journal *Metaphor and Symbol* (http://www.tandf.co.uk/journals/titles/1092-6488.asp), being published from 1986.

since the meaning of many LF values may keep constant (at least in the consciousness of speakers) when bases become abstract. Certainly, this property cannot be extended to all bases, and it is also subject to large cross-linguistic variation. Interestingly, the observation that paradigms of bases and collocates are much more restricted in their abstract, non-literal interpretations is found in early studies of collocations. Here is a clear example:

> "Whereas […] *explote* and *bomb* exhibit openness of collocability in relation to each other, *explode* in its figurative sense has a very limited collocational range: *myth, belief, idea, notion, theory*" (Cowie, 1981: 226).

As a matter of fact, this property of figurative interpretations, as regards "collocational ranges", holds generally. Consider the following LF in Spanish: $Oper_1$(*crisis*) = *atravesar*. The Spanish verb *atravesar* ('cross, go through') admits any direct object denoting a physical object or a place, but it is very restricted in its figurative interpretation. In this reading, it typically combines with nouns denoting situations of adversity: *crisis* 'crisis', *problema* 'problem', *dificultad* 'difficulty', *bache* 'bad time', *depresión* '*depression*', *vaivén* 'change of fortune, up and down', and a few other similar nouns belonging to a paradigm described in detail in the dictionary *REDES*. Again, it is rather unlikely that the Spanish verb *atravesar* in *atravesar el desierto* ('go through the dessert') has a meaning different to that of the same verb in *atravesar una depresión* ('go through a depression').

This aspect of the analysis of collocations is clearly related to the one discussed in 3.2. The LFs in the second column of the following table may be correctly formulated, but notice that it is almost impossible to relate the information displayed in the other columns without taking into account the speaker's knowledge of the collocate's meaning. We should also bear in mind the fact that abstract bases provide very restrictive paradigms in these cases, unquestionably related to the literal, physical or non-figurative meaning of the LF's values:

Collocate	Lexical function	Bases for the literal, physical or non-figurative interpretations of the LF's value	Bases for the abstract or figurative interpretations of the LF's value
torcerse 'bend, twist, go wrong'	$AntiFact_0$	*tobillo* 'ankle', *árbol* 'tree', *cable* 'cable'	*plan* 'plan', *proyecto* 'project', *previsión* 'forecast'
canalizar 'canalize, channel, carry'	$CausFact_0$	*agua* 'water', *río* 'river'	*ayuda* 'help', *sentimiento* 'feeling'
disolverse 'dissolve'	$FinFunc_0$	*azúcar* 'sugar', *polvo* 'dust'	*pacto* 'pact', *matrimonio* 'marriage'
cosechar 'harvest, win, achieve'	$Mult+Oper_1$	*trigo* 'wheet', *uva* 'grape'	*triunfo* 'victory', *beneficios* 'profits'
atravesar 'go through'	$Oper_1$	*desierto* 'desert', *muro* 'wall'	*crisis* 'crisis', *depresión* 'depression'

3.5 Lexical classes are recurrent

This is an important property of collocations. Since the lexical classes to which bases pertain are recurrent, reference to these semantic groups makes it unnecessary to multiply LFs for each member of the paradigm. For example, the following verbs of influence in Spanish

> *afectar* 'affect', *condicionar* 'condition', *gravitar* 'gravitate', *incidir* 'have an effect', *influenciar* 'influence', *influir* 'influence', *marcar* 'mark, determine', *pesar* 'weigh, carry weigh', *repercutir* 'affect, have repercussions',

are appropriate bases for the following adverbial collocates:

> *considerablemente* 'considerably', *decisivamente* 'decisively', *favorablemente* 'favourably', *inevitablemente* 'inevitably', *irremediablemente* 'irremediably', etc.

Certainly, we may find cases in which not all the members of a semantic paradigm of bases are entirely natural when combined with a particular collocate. The irregular combinations must be specifically marked in those cases, but the number of possible exceptions is insignificant as compared to the cost of the opposite option, that is, the task of specifying the full list of members of each semantic paradigm, instead of the appropriate semantic covering label.

In order to relate bases, collocates and the semantic notions connecting them, we used letters and numbers in *REDES*, since the dictionary is printed, and it is not available on line for the moment. Here is a fragment of an adverbial collocate entry:

decisivamente *adv.* ▮ Se combina con...

A VERBOS QUE DESIGNAN LA ACCIÓN DE TOMAR PARTE EN UNA ACTIVIDAD, MUY FRECUENTEMENTE UNA TAREA COMÚN A VARIAS PERSONAS: **1** contribuir ++: Secuestros en Colombia de 1995 y 1996, en la solución de las cuales Mauss contribuyó *decisivamente*. SEM210197 **2** colaborar ++: ...es un especialista en biomecánica de apoyos y hace algunos meses colaboró *decisivamente* en la recuperación del tenista... LVE120296 **3** intervenir ++: En las consecuencias finales de la movida monetaria intervendrán *decisivamente* muchos otros factores. LVE220495 **4** participar ++: ...en más de una ocasión ha participado *decisivamente* en desvelar casos que han estremecido a lectores y televidentes. CAP080198 **5** cooperar: ...la guerra de las galaxias puede enjuiciarse como un farol que cooperó *decisivamente* a que la URSS arrojara la toalla. LVE290996 **6** coadyuvar: Pero este viaje real ha coadyuvado *decisivamente* a obtener resultados favorables para la economía española. LVE241096 **7** tomar parte: ...una larga y enconada polémica en la que tomó parte *decisivamente* a lo largo de varios meses. INDOC **8** sumarse: *decisiva* por dar un nuevo paso. EPE191199 **25** incentivar: ...con ella se incentiva *decisivamente* a las pequeñas y medianas empresas... EME031296

D VERBOS QUE DENOTAN CAMBIO DE ESTADO O MODIFICACIÓN DE ALGO: **26** cambiar +: A veces, son viejos recortes los que cambian *decisivamente* la marcha de la trama. EME090494 **27** alterar +: ...en puridad analítica podrían estar casi un 10 por encima sin que la situación económica se alterase *decisivamente*. EME060294 **28** modificar: ...esquemas formales legados por la tradición escolástica, a los que modifican *decisivamente* la inventiva y pragmatismo de Chávez. PME210796 **29** variar: ...confidencias que nacen en el verano de 1991 y que hacen variar *de forma decisiva* la posición de Bono ante Guerra. EME200294

E VERBOS QUE DENOTAN AUMENTO O DESARROLLO. TAMBIÉN CON OTROS QUE DENOTAN MEJORA, PROGRESO O INCREMENTO DE ALGUNA MAGNITUD: **30** ampliar: Todos estos funestos episodios han sido descubiertos o ampliados *de modo decisivo* por El Mundo. EME031295 **31**

A series of semantic groups are displayed for each collocate. The bases exemplifying these groups contain marks of frequency (+, ++). They are followed by examples taken by a large press corpus. Literary sources were disregarded, since percentages of non-representative samples were proven to increase in significant ways in combinations extracted from literary texts.

Group B of this entry refers to verbs on influence (a fragment of this part of the entry follows). Notice that combination number 12 corresponds to the verb *afectar* 'affect':

B VERBOS QUE DENOTAN INFLUENCIA O REPERCUSIÓN: **10** influir ++: Mitterrand tiró el anillo al Sena y aquella ruptura debió influir *decisivamente* en su personalidad. EME090196 **11** pesar ++: El otro elemento que pesó *decisivamente* en el éxito aliado es que los alemanes nunca sospecharon seriamente de su máquina de cifrar. EME260295 **12** afectar ++: Hay otros factores, como las incertidumbres o el mercado de trabajo, que afectan *decisivamente* al cambio de actitud de las familias... LVE241095 **13** marcar ++: ...llegando a pasar por todas sus categorías en los doce años que marcaron *decisivamente* su carrera deportiva. ENH010201 **14** incidir +: ...su mayor producción no incide *decisivamente* en un mejoramiento económico para él, sino para el intermediario. LHG300497 **15** condicionar +: ...son la raíz profunda del deterioro académico y condicionan *decisivamente* cualquier reforma. PME221296 **16** influenciar: ...el no poder responsabilizar por la acción a un Gobierno que no puede influenciar *de manera decisiva*. EPE170977 **17** repercutir: ...las próximas semanas repercutirán *decisivamente* en la brillantez de la conclusión de su carrera profesional. EME130296 **18** gravitar −: La percepción de un mundo aún conflictuado, incierto e inseguro (...) sobre el que los Estados Unidos de América gravitan *decisivamente*... LNA080792

This piece of information is retrieved by the program, which displays the entry of the verbal base *afectar*:

afectar ♦ claramente, considerablemente[78], de cerca[37], decisivamente[12], de lleno, de pleno[25], desfavorablemente[2], directamente, especialmente, favorablemente[11], físicamente, fuertemente[24], gravemente[24], indefectiblemente[19], indirectamente, inmediatamente, intensamente, levemente, negativamente[1], notablemente[2], ostensiblemente[21], parcialmente, pasajeramente, peligrosamente[41], personalmente, poderosamente[6], positivamente, profundamente[74], proporcionalmente, psicológicamente, relativamente, sensiblemente, seriamente[2], severamente[35], significativamente, sustancialmente[45], tangencialmente, temporalmente, visiblemente

☐ Véase también: incidir, influir.

Non-numbered collocates in base entries were added manually, since the number of collocate entries in *REDES* is limited. Finally, the notion of influence, which corresponds to class B in this example, can also be targeted:

INFLUENCIA

♦ (SUSTANTIVOS) Véase: **abrumador[I], acatar[J], acusado[I], acusar[H], apreciable[F], atemperar[H], crucial[E], decisivo[H], decrecer[E], deducir[C], desmedido[I], determinante[C], dominante[B], emanar[E], fecundo[I], imborrable[C], insignificante[B], magnificar[B], ostensible[G], palpable[B], preponderante[B], profundo[F], rendirse (a/ante)[B], sacudir(se)[E], sobreponerse (a)[B], sopesar[D], sustraer(se) (de/a)[A], tangencial[H]**

♦ (VERBOS) Véase: **considerablemente[I], cordialmente[E], decisivamente[B], desfavorablemente[A], drásticamente[H], encarecidamente[B], enérgicamente[F], en mucho[E], favorablemente[C], fuertemente[E], gravemente[D], indefectiblemente[F], inevitablemente[E,J], inexorablemente[C], irremediablemente[G], ligeramente[F], maliciosamente[F], negativamente[A], notablemente[A], ostensiblemente[C], peligrosamente[H], plenamente[G], poderosamente[A], profundamente[J], severamente[F], sustancialmente[E], trágicamente[D], vagamente[D]**

☐ Véase también: ATRACCIÓN.

The information highlighted in the entry INFLUENCIA tells the user that in group B of the entry for the collocate *decisivamente*, he or she will find a paradigm of verbs of influence. Similar paradigms of verbs or nouns will be found for each reference in the list.

As can be seen, collocates are the focus of the lexical analysis in *REDES*. The semantic information provided in these entries is retrieved by a program in various ways, and presented to the user in the form of indexes.

3.6 Standard cases of semantic restrictions

Not much profit is taken in the literature on collocations from the well-known fact that predicates restrict the paradigms of their possible arguments. We may remember that the relevant semantic generalizations on predicate-argument traditional restrictions are often grounded in aspectual notions: verbs such as *witness, take place* or *recount* involve events; prepositions such as *during* take NPs denoting temporal units and also events, although perhaps not the same events that are possible in the complement of *on the occasion of* and other complex prepositions. On the other hand, a large number of well-known non-aspectual restrictions hold for verbs such as *play, wear, breath, nick* or *rescind,* to mention just a few, or adjectives such as *drastic, irrefutable, withering, effusive* or *crowded.*

Since most manner adverbs are predicates of events, parallel restrictions are expected for them on similar grounds: *overwhelmingly, impeccably, hectically, arduously, hermetically,* and many more. In fact, the attempt that I just made to provide possible bases for the adverbial collocate *substantially* was not different from the task of restricting the paradigm of possible

events selected by this predicate. This natural relation can be easily established if collocates are considered targets of research in themselves, rather than blind results of particular LFs.

3.7 Collocations vs. free word combinations

The deductive approach helps us to understand why the traditional opposition between collocations and free combinations is gradual. As a matter of fact, many of the so-called *free word combinations* are subject to abstract semantic restrictions which prove them not to be as free as they might look like. I have in mind expressions such as *Time flows,* or VPs such as *tell the truth; know English; close a door; supervise some work* or *win a match.* From a grammatical point of view, a verb restricts the paradigm of its possible objects in all these cases on the basis of some semantic features, but from a lexicographic point of view it is not absolutely clear how to tell whether or not these are collocations.

I must confess that I have always had more troubles than some of my colleagues telling apart collocations from free word combinations. If *write carefully* were a "free combination", as it is supposed to be, there would be nothing peculiar or strange in VPs such as *take a walk to the beach carefully, laugh carefully, imagine something carefully, kill someone carefully,* and many other VPs involving actions. I do not intend to describe the relevant generalization to be made on the grammatical properties of the adverb *carefully* (as it is obvious, a feature such as "action" would not do the work[2]), but simply point out that even the simplest candidates for "free word combinations" turn out to be part of restrictive semantic or pragmatic relations that we cannot hide or disregard. I am simply trying to emphasize that the problem with the standard term *free combination* is the exact meaning of the adjective *free* in this expression.

It is important to bear in mind that the existence of aspectual restrictions (i.e. based on the classes of *Aktionsarten*) on arguments by the predicates selecting for them —as opposed to restrictions based on other, more specific, semantic features— does not by any means imply that the sequences obtained stand for "free combinations". For example, the adverb *indefinitely* typically modifies either unbounded predicates (*continue, follow, hold, keep, occupy, persist, prevail, repeat, retain, seek, sustain, talk, wait,* etc.) or bounded predicates giving rise to resultative interpretations, and more frequently those which refer to the action of putting an end to something (*cancel, close, defer, delay, detain, isolate, open, postpone, put off, shut, suspend,* etc.). These restrictions are aspectual, rather than properly lexical, but it would be absurd to sustain that VPs containing this adverb are examples of "free word combinations".

The conclusion is straightforward. Many adverbs are predicates of events and turn out to be restricted on lexical or aspectual grounds. They do not seem to be collocates, but they do not give rise to free combinations either. In these and many other similar cases, the lexicographer might be happy to know that he or she is not facing a collocation, but the grammarian will be aware that, even so, there is likely to be some restriction to be accounted for. If, on the contrary, the lexicographer decides that he or she is meeting a collocation in some particular case,

[2] Lakoff (1973) is a classical piece as regards the various ways in which manner adverbs are lexically or pragmatically restricted.

the grammarian will equally reach the conclusion that there is some semantic restriction to formulate or some paradigm to characterize.

4 Some challenges of the deductive approach

4.1 The problem of relating semantic classes

Given a series of groups of bases for each collocate, a natural question arises: How are the several classes compatible with each collocate to be related to each other? At least two options present themselves:

1) *Attempt unification*. I have not been able to unify semantic classes (A, B, C…) in a single one for each collocate in most entries in *REDES*, but the attempt to reach unification to a certain extend is a quite reasonable enterprise. Moreover, this partial unification could provide crucial information for the task, by no means simple, of gathering verbs in semantic classes. Let me present an example of this. Here are two of the several lexical groups obtained for verbs which may be modified by the adverb *deeply*:

- *Breath, bury, delve, dig, embed, enmesh, fall, immerse, inhale, root, sigh, sink, …*
- *Analyze, examine, inquire, investigate, research, scrutinize, study, …*

Interestingly, verbs such as *penetrate* clearly belong to these two groups. It seems that unification will be a rather natural option here. It would tell us that *study*, *investigate* or *inquire* are interpreted linguistically as verbs denoting "movement towards the interior of something". This is a welcome result for a paradigm of verbs which do not seem to fit in other, more open, semantic classes, such as perception, communication, knowledge, etc.

I would like to emphasize the idea that focusing the meaning of collocates may lead us to a better understanding of the semantic classes to which bases objectively belong. In fact, the comparison of collocates is of great help in the task of classifying predicates. In *REDES* it is shown that the Spanish verb *leer* ('read') belongs to three different verb classes:

1. VERBS OF SPEECH: *leer en voz alta* 'aloud', *leer de carrerilla* 'non-stop', 'in one go', *leer atropelladamente* 'in a rushed way'.
2. VERBS OF PERCEPTION: *leer de refilón* 'slantingly', *leer entre líneas* 'between the lines', *leer por encima* 'leaf through, skim through', *leer de cerca* 'closely'.
3. VERBS OF CONSUMPTION: *leer ávidamente* 'eagerly, avidly', *leer con fruición* 'with relish', *leer vorazmente* 'voraciously', *leer compulsivamente* 'compulsively'.

Incidentally, no dictionary of Spanish seems to care about the third of these interpretations, a fact that might be of interest to lexicographers. Similar conclusions can be reached after a detailed comparison of similar collocates combining with other bases, whether verbal or not.

2) *Avoid unification.* Since it is not clear that lexical groups can always be unified, maybe they should be kept apart in some cases. For example, most verbal bases for the English adverb *substantially* denote changes of state, as we saw, but this semantic class does not seem to em-

brace verbs such as *aid, benefit, contribute, facilitate, help, support* and other similar to these, which may be modified by this adverb.

The policy of the dictionary *REDES* as regards this point was to keep lexical groups apart and hope that subsequent research will help us to reduce or integrate them. In more general terms, unification of lexical classes is always a tenable option, up to the point that the classes to be unified are not shown to play and independent role in other collocations. Again, this is an empirical issue, rather than a theoretical one.

4.2 Semantic vs. pragmatic paradigms

Any close scrutiny of the lexical paradigms that bases of collocations give rise to will reach the conclusion that the relevant information cannot always be formulated on semantic grounds. This is not a common situation, but it is attested in a number of cases.

Let us take some LFs such as $Oper_2$(*challenge*) = *face* or $Real_1$(*challenge*) = *face.* One might consider the issue whether or not one of them is more properly formulated than the other according to MTT principles, but I am more interested in another question, namely this: "What are the proper generalizations for nominal bases here"? We might try to isolate a group of nouns denoting situations of uncertainty (*challenge, choice, danger, dilemma, quandary, question, uncertainty,* …), another group of nouns denoting obstacles (*barrier, difficulty, hurdle, obstacle,* …), and perhaps a third one with nouns referring to nouns denoting situations of adversity (*death, catastrophe, loss, discrimination, drought, illness, punishment, pressure, prison, threat, violence,* …). These groups cover a large part of the possible nominal paradigm of bases that we might consider to be appropriate here, but this is not sufficient. The reason is, simply, that we can face the truth, face the world, face reality, face life or face the consequences of some action, and these nouns do not belong to a semantic paradigm. That is, the noun *barrier* intrinsically denotes an obstacle, but the noun *world* does not belong to a semantic paradigm of nouns denoting situations of difficulty or adversity.[3]

The solution that we gave to this problem in the project *REDES* was somehow similar to the standard solution given to conversational implicatures in Pragmatics: we may enlarge some paradigms (for example, those of nouns denoting adversity or difficulty) with items which do not belong to them from a lexical point of view (*world, truth, life,* etc.). This process is an attributive operation, since we predicate the features characterizing the paradigm (for example 'being difficult, adverse, hard, etc') of the new nominal elements that come to be integrated in it. A very natural question to be asked now is what lexical paradigms allow for this *pragmatic enlargement,* and which ones reject that possibility. This is another interesting, as well as promising, line of research that the project *REDES* suggests.

[3] The fact that paradigms in collocations may meet some pragmatic, rather than semantic, criteria is sometimes implicit in the literature. An example is Stubbs's (1996) observation that the complements of the verb *cause* typically refer to negative notions (*accident, damage, death,* etc.) whereas complements of *provide* typically denote positive notions (*care, shelter, food,* etc.).

5 Conclusion: Back to controversial features of collocations

In the outset of this paper, I distinguished non-controversial features of collocations from controversial features, and I pointed out that the latter seem to be much more interesting than the former. The deductive approach to collocations that I have sketched here, and developed in the dictionary *REDES*, points towards the following evaluation of controversial features:

1. We may accept that collocations are salient, but we have to be sure that saliency is not a merely statistical notion.

2. Most collocations are not binary, since bases form large paradigms that can be defined and characterized on semantic grounds. If it is true that a language might contain around half a million collocations, as it is often argued, a straightforward conclusion follows: collocations cannot be memorized by speakers, whether native or not. In the deductive approach to collocations that I favor, it is assumed that speakers have access to the semantic notions involved in paradigms of bases, and also that collocates are sensitive to these abstract semantic features. Most of these paradigms are open, and a few of them might even be characterized on pragmatic (i.e. not just semantic) grounds.

3. Most collocations are not arbitrary, since the fact that some particular lexical items belongs to some semantic class is not an arbitrary fact, but rather a direct consequence of its meaning. Certain collocations (for example, those constituted by light verbs) seem to present a higher degree of arbitrariness, but even in these cases a number of generalizations are possible (as shown in Sanromán Vilas, 2011).

4. There is no simple answer to the question whether or not collocations are transparent relations. I believe that they are, to a certain extent, as a natural conclusion of the fact that collocates are meaningful.

Acknowledgements

Many thanks to Auxi Barrios for helping me to formulate Lexical Functions according to MTT standards. Needless to say, all possible errors or shortcomings are my own.

Bibliography

A. Dictionaries

BBI: M. Benson, E. Benson & R. Ilson. *The BBI Combinatory Dictionary of English. A Guide to Words Combinations,* Amsterdam: John Benjamins, 1986.

CCEC: *Collins Cobuild English Collocations on CD ROM.* New York: Harper Collins Publishers, 1995.

MCD: *Macmillan Collocations Dictionary for Learners of English*, coordinated by M. Rundell. Oxford: Macmillan, 2010.

OCD: *Oxford Collocations Dictionary for Students of English*, Oxford: Oxford University Press, 2002.

REDES: *Redes. Diccionario combinatorio del español contemporáneo*, coordinated by I. Bosque. Madrid: SM, 2004.

B. Other references

Aisenstadt, E. 1979. Collocability Restrictions in Dictionaries. In Hartmann, R. (ed.). *Dictionaries and their Users.* Exeter Linguistic Studies 4, 71-74.University of Exeter.

Apresjan, J. 2009. The Theory of Lexical Functions: An Update. In *Proceedings of the Fourth International Conference on MTT*. 1-14. Montreal: OLST. http://meaningtext.net/mtt2009/01-JApresjan.pdf

Apresjan, J. & M. Glovinskaja. 2007. Two Projects: English ECD and Russian Production Dictionary. In *Proceedings of the 3rd International Conference on MTT.* München: WienerSlawistischer Almanach. Available at: http://meaningtext.net/mtt2007/proceedings/03Apresjan_GlovinskajaFinal.pdf

Bertinetto, P. M. & M. Squartini, 1995. An Attempt at Defining the Class of Gradual Completion Verbs. In Bertinetto, P. M., et. alii (eds.). *Temporal Reference, Aspect and Actionality. Vol. 1, Semantic and Syntactic Perspectives*, 11-26. Torino: Rosenberg & Sellier.

Bosque, I. 2001a. Sobre el concepto de 'colocación' y sus límites, *Lingüística Española Actual*, 23(1): 9-40.

Bosque, I. 2001b. Bases para un diccionario de restricciones léxicas, *Moenia* 7:11-52.

Bosque, I. 2004a. Combinatoria y significación. Algunas reflexiones. In Bosque, I. (dir.), *Redes. Diccionario combinatorio del español contemporáneo,* lxxvii-clxxiv. Madrid: SM.

Bosque, I. 2004b. La direccionalidad en los diccionarios combinatorios y el problema de la selección léxica. In Cabré, T. (ed.). *Lingüística Teòrica: Anàlisi i perspectives I,* Catalan Journal of Linguistics Monographies, 3-58. Available at: http://filcat.uab.es/clt/publicacions/coleccions/monografies/pdf/LT-I-Bosque.pdf

Cowie, A. P. 1978. The Place of Illustrative Material and Collocations in the Design of a Learner's Dictionary. In Strevens, P. (ed.). *In Honour of A.S. Hornbyk,* 127-139. Oxford: Oxford University Press.

Cowie, A. P. 1981. The Treatment of Collocations and Idioms in a Learner's Dictionary, *Applied Linguistics* 2(3): 223-35.

Cowie, A. P. 1983. General Introduction. In Cowie, A. P. et al. (eds.). *Oxford Dictionary of Current Idiomatic English*, volume 2, 10-17. Oxford: Oxford University Press.

DeCesaris, J. & P. Battaner, 2006. A New Kind of Dictionary: *REDES. Diccionario combinatorio del español contemporáneo*. In Corino, E. et al. (eds.). *Proceedings XII EURALEX In-*

ternational Congress, 399-408. Alessandria: Edizioni dell'Orso. Available at: http://www.euralex.org/elx_proceedings/Euralex2006/050_2006_V1_Janet%20DECESARIS,%20Paz%20BATTANER_A%20New%20Kind%20of%20Dictionary_REDES,%20Diccionario%20combinatorio.pdf

Declerck, R. 1979. Aspect and the Bounded/Unbounded (Telic/Atelic) Distinction, *Linguistics* 17:761-794.

Fauconnier, G. & M. Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*, New York: Basic Books.

Glucksberg, Sam. 2001. *Understanding Figurative Language*, Oxford; Oxford University Press.

Hausmann, F. J. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen, *Praxis des Neusprachlichen Unterrichts*, 31:395-406.

Hausmann, F. J. 1989. Le dictionnaire de collocations. En F. J. Hausmann et alii (eds.), Wörterbücher/Dictionaries/Dictionnaires, vol. 1, Berlin y New York, Walter de Gruyter, págs. 1010-1019.

Kennedy, C. & B. Levin. 2008. Measure of Change: The Adjectival Core of Degree Achievements. In McNally, L. & C. Kennedy, (eds.). *Adjectives and Adverbs: Syntax, Semantics and Discourse*, 156-182. Oxford: Oxford University Press.

Lakoff, G. 1973. Notes on what it would take to understand how an Adverb works, *The Monist* 57: 328-343.

Levin, B. & M. Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Semantics Interface.* Cambridge, MA: MIT Press.

Mel'čuk, I. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Wanner, L. (ed.). *Lexical Functions in Lexicography and Natural Language Processing*, 37-102. Amsterdam: John Benjamins,

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Sanromán Vilas, B. 2001. The Unbearable Lightness of Light Verbs. Are They semantically Empty Verbs? In this volume, 253-263.

Stubbs, M. 1996. *Text and Corpus Analysis*, Oxford: Blackwell.

Stubbs, M. 2001. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Tenny, C. 1994. *Aspectual Roles and the Syntax-Semantics Interface*, Dordrecht: Kluwer Academic Publishers.

Vega Moreno, R. 2007. *Creativity and Convention: The Pragmatics of Everyday Figurative Speech*, Amsterdam: John Benjamins.

Collocations in Basque: a test for classification[1]

Xabier Altzibar, Xabier Bilbao, Koldo Garai

Department of Basque Philology – University of the Basque Country
josejavier.alcibar@ehu.es, xabier.bilbao@ehu.es, koldo.garai@ehu.es

Abstract

This work presents an initial attempt to classify collocations in Basque, since there are at present no dictionaries of collocations or word combinations in this language. We take as our reference point the taxonomy of Spanish by G. Corpas (1996), making allowances for the distinctive features of Basque morphosyntactic structures and current usage. We classify the collocations by their morphosyntactic structure and specify the base in each case, all with examples of common productions taken from the media and from spoken language. We also establish the principal lexical functions that appear in the examples.

Keywords

Lexicology, Phraseology, Collocations, Word combinations, Basque

1 Introduction

Our objective in this paper is to suggest a classification of Basque collocations by their base and morphosyntactic structure. For this purpose, we have collected and examined a number of common examples from the media and spoken language. There is no unitary definition of collocation, but two main approaches can be distinguished: one views collocation on the basis of the frequency with which two lexical units co-occur; the second views collocations as being the result of giving lexical form to a number of semantic relations that can be formally described (Bosque (dir.) 2004:CLIII). We take as our base the taxonomy of Spanish by G. Corpas (1996:66-76), although we feel allowances need to be made for the specifics of Basque grammatical categories. We seek to map the different types of Basque collocations in the grammar. In addition, many of the examples are classed into different lexical functions (Mel'čuk, 1998), although our main purpose is not to offer an exhaustive application of

[1] This article is a product of an ongoing project by the research group EHU10/19 supported by the University of the Basque Country (UPV-EHU). We would like to acknowledge the abundant support and advice we have received from our group co-members Juan Carlos Odriozola, Jose Ramon Etxebarria and Ainara Ondarra. We would also like to thank the three anonymous Barcelona-MTT reviewers of the first version for their helpful comments and suggestions.

lexical functions. In fact, we believe that this approach may be useful for compiling general or specialised dictionaries and for teaching Basque as a second language. A taxonomy of Basque collocations would also be useful for the typology of languages. Following we provide a quick glance of Basque's typology and its social context as a backgroud for the analysis this paper presents.

Basque is an agglutinative language; unlike prepositive languages such as Spanish, French or English, it operates with cases and postpositions attached to the noun. Examples of cases are absolutive, ergative and dative, thus: *etxe* 'house' + *a* (determiner of absolutive case): *etxe**a*** ['house-abs][2] 'the house'; *etxea**k*** or ergative case, [house-erg] 'the house'; *etxea**ri*** or dative case [house-dat] 'to the house'. Examples of postpositions attached to the noun are, *inter alia*, allative, inesive, genitive-locative and instrumental, thus: *etxe**ra*** or allative [house-all] 'to the house'; *etxea**n*** or inesive [house-in] 'in the house'; *etxea**z*** or instrumental [house-ins] 'by (means of) the house'. Genitive-locative as in *etxe**ko***, [house-gen.loc] 'of the house, is a multifunctional postposition which can function as an adjective e.g. *erabateko akordio* 'total agreement' and as a nominalising suffix, e.g. *ezusteko galanta* [surprise great-abs] 'a great surprise'.

Basque is currently undergoing a process of unification, standardisation, and adaptation to new uses. However, the influence of Spanish and French, especially through calque, is having an almost decisive influence on this process of renovation (Alberdi et al., 2010). Indeed, over the last forty years a host of new collocations has entered the language of the Basque media (Altzibar, 2005:383-395). This paper is an attempt at a classification of collocations in Basque, a field that is arousing increasing interest and inspiring ever more research in Basque, such as the current work on phraseology by members of the lexicographically-oriented Elhuyar Fundazioa. Although Basque dictionaries and corpuses provide much information, there are still no dictionaries of collocations, and we have therefore had to identify many of them through observation and intuition. The examples —we have chosen only a few and left out the many dialectical and other variations— are taken from the daily newspaper *Berria* and from some dictionaries and compilations of popular lexicon.

We shall classify the collocations into three classes, based on their morphosyntactic structure: noun-based, adjective-based, and adverb-based.

2 Noun-based collocations

Basque noun phrases may be of three subclasses: when they function as the subject of a transitive verb they take the ergative case (*-k* added directly to the word or to the end of the determiner); for the indirect object the phrase takes the dative case (*-(r)i* added to the word). The absence of any addition, with the word appearing unadorned or with the determiner (*-a* sing., *-ak* plur.) alone, indicates the absolutive case. This is the most difficult situation, since the absolutive operates as an object with transitive verbs, but as a subject with intransitive verbs. This poses a problem for our classification because, except for clear examples, an

[2] Literal representations of the Basque examples are provided between brackets. A hyphen separates case suffixes and postpositions from the stem; a dot separates compounded stems and/or derivational suffixes. Abbreviations used are as follows: abs: absolutive; abl: ablative; all: allative; dat: dative; det: determiner; erg: ergative; gen: genitive; ins: instrumental; LF: lexical function; loc: locative; sin: singular; pl: plural; aux: auxiliary; tran: transitive; intr: intransitive.

object might function as a subject if used in an impersonal sentence (e.g. z*urrumurrua zabaldu da* [rumour-abs spread has] 'The rumour (has) spread' vs. *Irratiak zurrumurrua zabaldu du* [radio-erg rumour-abs spread has] 'The radio (has) spread the rumour': the absolutive z*urrumurrua* functions as an object in the latter example, and as a subject in the former). The absolutive case is therefore in itself functionally ambiguous. We shall begin our description by presenting the unambiguous cases: the ergative (always subject) and the dative (always indirect object except for defective verbs).

2.1 Noun (ERG) + verb

The subject receives the *-k* marker of the ergative or active subject:

> *etsiak hartu* [hopelessness-erg take] 'to be overcome by hopelessness'; *(mendia) suak hartu* '[(mountain-abs) fire-erg take] 'the fire catch(es) (the hill)'; *loak hartu* [sleep-erg take(s)] 'to fall asleep'; *goseak hil* (in examples such as *goseak hiltzen nago, I'm dying of hunger' fig.*) [hunger-erg kill] 'the hunger kill(s)'; *janariak/edariak on egin* [food/drink-erg well done] 'the food/drink do(es) good.'

The examples above are of the value Oper '(to) do, carry out'. Non-ergative versions of some of those collocations can be found, as in *etsia hartu, su(a) hartu, lo(a) hartu* and *gosea hil*.

2.2 Noun (DAT) + verb

Collocations with the noun in the dative case, *-(r)i*. Here only examples where the dative is required by the verb are presented, without necessarily implying an indirect object function.

> *temari eutsi* [obstinacy-dat hold] 'to insist on (sth.)'; lexical solidarity exists between the two words (LF=ContOper); *lanari ekin* [work-dat insist] 'to apply (oneself) to work' (LF=IncepReal$_1$); *ihesari eman* [escape-dat give] 'to escape/flee' (LF=IncepOper).

2.3 Noun (absolutive case = subject/object) + verb

Contextual information is normally crucial in determining the function (subject or object of the absolutive noun), i.e. in few cases is the perceived function "context-free", and these are mostly instances when the absolutive functions as an object.

2.3.1 Noun (absolutive = object) + verb

This is a very frequent subtype and includes collocations with a wide range of combining possibilities: e.g. in some units the collocate (verb) can be used with different noun bases from the same semantic field, as in *eztabaida/auzia erabaki* [argument/dispute-abs resolve] 'to resolve an argument/dispute'. Other examples:

> *garaipena lortu* [victory-abs achieve] 'to achieve the victory'; *gatazka konpondu/gainditu* [conflict-abs repair/overcome] 'to resolve the conflict'; *kargua izan/bete/utzi* [post-abs have/fulfil/leave] 'to hold/fulfil/leave the position (of responsibility)'; *kontrola eduki/galdu* [control-abs have/lose] 'to hold/lose control.'

The lexical functions of the examples above are $Oper_1$ (*garaipena lortu, kontrola eduki*), and $Real_1$ (*gatazka konpondu/gainditu, kargua bete*).

Within the noun (abs=object) + verb type, probably the most numerous collocations are those in which the verb is delexicalized and its meaning nearly grammaticalized, with the noun providing the semantic content. These are polysemic verbs: *eman* 'give', *hartu* 'take', *ipini/jarri* 'put/place', *ekarri* 'bring'...

> *hasiera/amaiera eman* [start/end-abs give] 'to commence'; *erabakia hartu* [decission-abs take] 'to make a decision'; *mozorroa ipini* [*disguise-abs* put/place] 'put on a disguise.'

The examples above bear the lexical functions $Oper_1$ (*erabakia hartu),* $CausFact_0$ (*hasiera/amaiera eman),* $Real_1$ 'real' (*mozorroa ipini*).

This class contains a large subtype, which includes the collocations formed with almost completely delexicalized verbs: *egin* 'to do', *izan* 'to be', *ukan* or *eduki* 'to have/possess, *hartu* 'take', as in:

> *lotsa izan* [shame-abs have] 'to feel ashamed'; *parte hartu* [part-abs take] 'to take part'; *arrakasta izan/ukan/eduki* [success-abs-sin have] 'to be successful'; *hitz egin* [word-abs make] 'to speak.'

These examples are of the value $Oper_1$. The most singular verb collocate in Basque is *egin* 'to make/do': in Basque common actions such as *speak, sleep and work* are not expressed by means of single-word verbs, but by noun + verb pairs: *hitz* 'word', *lo* 'sleep', *lan* 'work' + verb *egin* 'make'. This is a very productive device in Basque, and is used to form many verbs (see Basque constructions with support or light verb in Alonso Ramos 2001:88-93; Abaitua 1988).

2.3.2 Noun (ABS=subject/object) + verb

Here we present some examples in which the function of the absolutive may vary depending on whether the verb is intransitive (impersonal sentences) or transitive. In the latter case the absolutive will clearly be the object of the verb:

> *gerra piztu (da)* [war-abs lit (aux-intransitive)] 'the war (is) started' (there is an underlying metaphor: "war is fire"); compare with *gerra piztu (du)* [war-abs lit (aux-transitive)] '(somebody) has started the war'; *susmoa egiaztatu (da)* [suspicion-abs confirm (aux-intransitive)] 'the suspicion (is) confirmed'; compare with *susmoa egiaztatu (du)* [suspicion-abs confirm (aux-transitive)] '(somebody) (has) confirmed the suspicion'; *zurrumurrua zabaldu (da)* [rumor-abs spread (aux-intransitive)] 'the rumour (is) spread'; compare with

zurrumurrua zabaldu (du) [rumor-abs spread (aux-transitive)] '(somebody) has spread the rumour.'

These examples bear the lexical functions $Func_0$ 'to function', e.g. *zurrumurrua zabaldu (da)*, CausFunc, e.g. *zurrumurrua zabaldu (du)*, *gerra piztu (du)*, IncepFunc, e.g. *gerra piztu (da*), and Fact 'fact' (e.g. *susmoa egiaztatu*).

2.4 Noun + adjective

In many collocations, the collocate (adjective) intensifies its base (noun). The lexical function Magn 'very', 'to a (very) high degree', 'intense(ly)', 'completely' is one of the most frequent, e.g. *adiskide/lagun zahar* [friend/companion old] 'old friend'. The collocate *min* 'sharp, intense', is combined with various nouns:

adiskide min [friend intense] 'close friend*; etsai min* [enemy intense] 'bitter enemy'*; dolu min* [pain sharp] 'deep pain' (lexicalised in *dolumin* 'condolence')*; uda min* [summer intense] 'the height of summer'*; negu min* [winter intense] 'the critical period of winter.'

Other adjectives (*hotz, huts*) collocate with specific nouns. With *hotz* 'cold; simple, mere, pure':

haur hotz [child mere] 'a mere child'; *hil hotz* [dead cold] 'dead body'; in Basque, this pair has been lexicalized as 'corpse'; *pobre hotz* [poor mere] 'utterly poor'*; alfer hotz* [lazy pure] 'utterly lazy' (these last two examples can also be considered as instances of 'adjective + adjective' collocations; see. 3.1).

Similar to *hotz* but semantically broader is *huts* 'mere, pure, simple; alone':

logika hutsa [logic alone] 'pure logic'*; haur huts* [child mere] 'a mere child'; *ganorabako huts* [foundation.without pure] 'utterly irresponsible (person).'

Some very common values of the lexical function Magn are interchangeable when they accompany some nouns: *bizi* 'living, lively', *larri* 'serious, acute', *gorri* 'red, raw', *gogor* 'hard', *amorratu* 'furious', *porrokatu* 'waste', etc.:

istilu larri/gorri/gogor [conflict acute/raw/strong] 'serious conflict'*; haserre/eskandalu bizi/gorri* [indignation/scandal live/raw] 'live indignation/scandal.'

But the very same adjectives are not interchangeable with other nouns:

borroka bizi [fight live] 'violent fight'; *gogo bizi* [desire live] 'living desire', whilst there is no **gogo gorri* [desire raw] 'strong desire', or **gogo larri* [desire acute] 'pressing desire', or **gogo gogor* [desire hard] 'strong desire.'

Collocations with polysemous colour epithets (*gorri 'red', beltz* 'black', *zuri* 'white') are abundant in Basque. Those with lexical function Magn include*:*

> *miseria gorri* [misery red] ‘extreme misery’*; behar/premia gorri* [need red] ‘pressing need’*; istilu gorri* [conflict red] ‘grave conflict’*; eskandalu gorri* [scandal red] ‘grave scandal’; *miseria gorri* [misery red] ‘extreme destitution’; *ele zuri* [word white] ‘false word.’

Also abundant are collocations of lexical function Epit ‘redundant cliche’, i.e. of a semantically empty epithet:

> *ezpain gorri* [lip red] ‘lip’*; urre gorri* [gold red] ‘gold’; *zilar zuri* [silver white] ‘silver’; *gau beltz* [night black] ‘dark night’; *itsaso zabal* [sea wide] ‘vast/immense ocean/sea’; *mundu zabal* [world wide] ‘wide world.’

Collocations of lexical function Ver ‘real’, ‘genuine’, ‘as it should be’: *labana zorrotz* [knife sharp] ‘sharp knife.’

In some examples of this LF the same collocate (adjective) is shared by various (base) nouns from the same semantic domain:

> *bista/begi/belarri zorrotz* [sight/eye/ear sharp] ‘sharp sight/eye/hearing’; *ahots/begi/belarri zoli* [voice/eye/ear fine] ‘fine voice/eye/hearing.’

Many of the collocations of lexical function Ver are exemplary or ideal-confirming:

> *euskaldun garbi* [Basque clean] ‘pure Basque’; *euskaldun peto-peto* [Basque fully] ‘thorough Basque’; *euskaldun jator* [Basque authentic] ‘authentic Basque.’

Some fixed descriptions can be classed within the lexical function Bon ‘good’. Traditional examples include:

> *Jainko on* [God good] ‘the good God’*; Jainko txar* [God bad] ‘the bad God’ does not exist; *patxada on/eder* [calm good/plentiful] ‘good/great parsimony’*; sasoi on/eder* [age/season optimal] ‘optimal time (esp. about physical condition), time of youth.’

Others have a collocate with a negative significance (LF AntiBon):

> *de(a)bru (t)zar* [devil bad] ‘the bad devil’; *jenio txar/gaizto* [temper bad/evil] ‘bad temper’; *susmo txar* [suspicion bad] ‘bad suspicion.’

There are abundant collocations of two opposite variants (positive/negative, good/bad, correct/incorrect, right/wrong), of the value Pos ‘positive evaluation’ and AntiPos.

> *balantza positibo/negatibo* [balance positive/negative] ‘positive/negative balance’; *egoera on/txar* [situation good/bad] ‘good/bad situation’; ohitura *on/gaizto* [habit bad] ‘bad habit’; *iritzi on/txar* [opinion good/bad] ‘good/bad opinion’*; borondate on/txar* [will bad/good] ‘good/bad will.’

2.5 <adjective suffix -ko> + noun

Most adjectives follow the noun in Basque, but adjectives formed by the derivational suffixes *-dun, -ko* usually precede the noun. There are a number of adjective + noun pairs where the *-ko* adjective acts as a collocate:

> *aldeko/kontrako iritzi/boto* [favour.of/against.of opinion/vote] 'favourable/opposing opinion/vote'; *erabateko akordio* [total.of agreement] 'total agreement'; *erabateko lehentasun* [total.of priority] 'total priority.'

In these examples, the main lexical functions are Pos (*aldeko iritzi, boto*), AntiPos (*kontrako iritzi/boto*), and Magn (*erabateko akordio/lehentasun*).

Some of these *-ko* adjectives can be found functioning as nouns, and could therefore be classified under the noun + adjective class, as in:

> *ezezko borobil (eman/hartu)* [negative.of round (give/receive)] '(to give/receive) a round negative'; *ezusteko handia/galanta (izan)* [unexpected.of big/monumental] 'great surprise (to be).'

2.6 Noun + noun

In these collocations, the word order in Basque is opposite to Spanish; in English both order types are acceptable: the Spanish <noun + *de* + noun> constructions (*casa de muñecas* 'dolls' house', *ciudad de alegría* 'city of joy') have their counterpart in the Basque <noun + noun> constructions, where the order of the nouns is the reverse of Spanish: *panpin etxea* (lit. doll house), *alaitasun herria* (lit. joy city) 'city of joy'. Both constructions are possible in English, as the examples show. This <noun + noun> construction gives rise to a number of collocations, which can be divided into two lexical functions (Sing and Mult):

> Collocations bearing the lexical function Sing 'singular, unique':
>
> *baratxuri atal* [garlic part] 'clove of garlic'; *azukre koskor* [sugar piece] 'sugar lump'; *txokolate tableta/pastilla* [chocolate tablet/bar] 'chocolate bar'; *xaboi pastilla* [soap bar] 'bar of soap.'
>
> Collocations bearing the function Mult 'multitude':
>
> *arrain talde* [fish group] 'school of fish'; *erle-mulko* [bee-bunch] 'swarm of bees'; *txori talde/aldra/saldo* [bird group/band/flock] 'flock of birds.'

3 Adjective-based collocations

3.1 Adjective + adjective

Some of these double-adjective collocations can function as nouns, as in the first example

below (*aberats okitu*). The examples we present here bear the lexical function Magn and they have as their collocates the adjectives *okitu* 'fed up, full', *erromes* 'poor, pilgrim', *hotz* 'cold; simple, mere, pure', *huts* 'mere, pure, simple; alone', *garbi* 'pure, manifest', *arrail* 'long thick splinter; drunk', *amorratu* 'enraged':

> *aberats okitu* [rich full] 'very rich, millionaire'; *zahar okitu* [old full] 'very old' (but not **pobre okitu* [poor full]); *pobre erromes* [poor poor/pilgrim] 'very poor'; *pobre hotz* [poor pure] 'utterly poor'; *alfer hotz* [lazy pure] 'utterly lazy'; *lapur garbi* [thief pure] 'very thieving' (but not **aberats garbi* [rich pure]); *txarri zikin* [pig dirty] or *txerri urde* [pig filthy/hog] 'filthy pig'; *asto ziri* [donkey stupid] 'stupid donkey.'

The adjective *arrail* is more restrictive: it collocates only with *mozkor*/*hordi* 'drunk': *mozkor/hordi arrail/arraildu* [drunk split] 'completely drunk.'

These units also have another construction and pronunciation (with a pause in the middle): *aberatsa, okitua* [rich-abs, full-abs] 'very rich'; *zaharra, okitua* [old-abs, full-abs] 'very old'; *pobrea, erromesa* [poor-abs, poor/pilgrim-abs] 'very poor'; *alferra, hotza* [lazy-abs, pure-abs] 'utterly lazy'; *lapurra, garbia!* [thief-abs, pure-abs] 'very thieving'; *mozkorra, arraila* [drunk-abs, split-abs] 'completely inebriated.'

3.2 Adverb of degree formed with Ø or *-ki* + participle / adjective

The base of these units is the adjective or the participle in adjectival function, and the adverb is a collocate: *erabat komentzituta/komentzitua* [utterly convinced], *larri(ki)/arin(ki) zaurituta/zauritua* [grave(ly)/slight(ly) wounded], *zabal idekia* [wide open] (these examples are related to the adverbial constructions in 4.1.1). Adverbial suffix *-ki*, used in northern dialects, is optional. In the oxymoron *Itsuski ederra* [horribly beautiful] 'very beautiful' there is lexical solidarity between the collocates. All these examples implement the lexical function Magn, except *arin(ki) zaurituta/zauritua*, which is AntiMagn.

4 Adverb (adverbial phrase) + Verb

We have divided these adverb-based collocations into four classes: a first one with adverbs that characterize the manner and place of the verb's action; a second one with adverbs that indicate semantic and phonic proximity; a third one with onomatopoeia in adverbial function and a fourth where the adverb is a phrase based on a noun plus the inesive, allative or instrumental postpositions.

It is a matter of some controversy whether the adverb or the verb should be considered as the base of the collocation. In some examples, the verb seems to clearly merit being accepted as the base (e.g. *estu lotu* [tight tie] 'to tie tightly'), but many other examples are much harder to decide upon, especially in the subtypes Adverb (semantic and phonic proximity) + Verb, Onomatopoeia (in adverbial function) + Verb, or Adverbial phrase + Verb, as exemplified below.

The most frequent adverb suffix is Ø; other adverbial suffixes are *-ki, -ka* (derivational suffixes of manner), and the postpositive suffixes *-n* (inesive), *-ra* (allative), *-z* (instrumental).

4.1 Adverb (manner / place) + Verb

4.1.1 Adverb (manner) + verb (these examples bear the function Magn, except arin(ki) zauritu, which bears LF AntiMagn)**:**

> *argi (eta garbi) esan* [clear (and clean) say] 'to say clearly'; *arlote jantzi* [beggar dress] 'to dress like a ragged person'; *estu lotu* [tightly tie] 'to tie tightly'; *itsutuki maite* [blind.ly love] 'to love blindly'; *tinko/irmo eutsi* [firmly/steadfastly maintain] 'to maintain firmly'; *larri(ki)/arin(ki) zauritu* [grave(ly)/light(ly) injure] 'to be seriously/slightly injured'; *zabal ireki* [wide open] 'to open wide.'

Some of the adverbial collocates of manner have alternate antithetical adverbs (*ongi/gaizki* 'well/badly', *zuzen/oker* 'correctly/incorrectly', *alde/kontra* 'in favour /against'): *ongi/gaizki erabili* [right/wrongly treat) 'to treat well/badly'; *ongi/gaizki egin* [well/badly do] 'to do/act well/badly'. Examples:

> *ongi/gaizki esan (norbaiti)* [well/badly say (somebody-dat)] 'to speak kindly/unkindly (to a person)'; *zuzen/oker jokatu* [right/wrongly act] 'to act correctly/incorrectly, wrongly'; *alde/kontra izan/jarri/jokatu* [in favour/against be/place/act] 'to be/place (oneself)/to act in favour / against.'
>
> These examples bear the lexical functions Pos and AntiPos.

4.1.2 Adverb (place) + verb

There are also abundant examples of adverbial collocate + delexicalized verb (*egon* 'to be', *gelditu* 'to stay', *ibili* 'to walk' etc.):

> *kanpo(an)/barne(an) izan/egon* [outside(-in)/inside(-in) be/stay] 'to be outside/inside' (LF Loc_{in}); *bazter(ean) gelditu/egon/utzi* [side(-in) remain/stay/leave] 'to remain/be/leave aside' (LF Loc_{in}); *aurrera egin* [forward-all do] 'to go forward, advance' (LF Loc_{ad}); *atzetik ibili* [back-abl move] 'to go behind, follow' (LF Loc_{ab}).

4.2 Adverb (semantic and phonic proximity) + Verb

Somewhere on the border between collocations and idioms are units with a double and (quasi)synonymic collocate. Some of these dual collocates also implement the lexical function Magn:

> *negar zotinka ari/egon* [cry sobbing act/be] 'to be sobbing'; *antsika eta negarrez ari/egon* [moaning and crying act/be] 'to be moaning and crying'; *negar marrumaz ari /egon* [cry howling act/be] 'to be sobbing and shouting.'

As the examples above show, in some of these expressions the two near-synonymic collocates are conjoined by the conjunction *eta* 'and' —a structure with an intensive reduplicative function:

> *korrika eta presaka (joan/ibili)* [running and rushing (go/act)] '(go/act) in a hurry'*; estu eta larri (ibili) [pressed and gravely (go/act)]* '(to act/go) very hurriedly.'

Sometimes the semantic proximity mentioned above may be reinforced by some phonic devices, such as rhyming pairs:

> *sor eta gor egon/gelditu* [perceptionless and deaf be/remain] 'to be/remain completely deaf', lexicalized in *sorgor* 'deaf, insensitive'; *sor eta lor gelditu* [perceptionless and upset remain] 'to become completely astonished'; *isil eta biribil egon* [quiet and round be] 'to be/remain silent and fulfilled/entire.'

Among these, there are collocations formed by two adjectives with different linguistic origins (with an attributive or predicative adverbial function), one from Basque and another from Spanish, both of them bearing a close or similar meaning; the play between the languages lends expressivity, humour and colloquial connotations to these units, which are often used without a verb:

> *antzeko parezido (izan)* [similar alike (be)] '(to be) similar'

Sometimes both words come from Spanish directly, keeping an assonance or loose rhyme between them, as in:

> *tentepotente (egon)* [upright.strong (be)] '(to be) standing, still/stiff'; *serio demonio egon/ari* [serious devil be/act] '(to be/to act) seriously'; *kieto para(d)o (egon)* [still still (be)] '(to be) stock still.'

4.3 Onomatopoeia (in adverbial function) + verb

In Basque, collocations formed by onomatopoeias are frequent in both common and educated language.

> *dinbi-danba jo* [<symbolic sound> hit] 'to hit repeatedly, to fire a shot, to toll (a bell)'; *zanga-zanga edan* [<symbolic sound> drink] 'to drink with large gulps or with great desire'; *tipi-tapa joan* [<symbolic sound> go] 'to march/walk step by step'*; mauka-mauka jan* [<symbolic sound> eat] 'eat voraciously'; *bala-bala/barra-barra zabaldu* [<symbolic sound> spread] 'to spread'; *(elurra) mara-mara ari/egin/erori/bota* [(snow-abs) <symbolic sound> do/make/fall/throw] 'to snow gently'; *(elurra) zarra-zarra ari* [(snow-abs) <symbolic sound> do] 'to snow heavily'*; (euria) ziri-ziri/ziri-miri egin/ari/erori/bota* [(rain-abs) <symbolic sound> make/do/fall/throw] 'to rain gently' (from *ziri-miri* comes the word *sirimiri* also used in Spanish for 'drizzle'), etc.

In some of these examples there is lexical solidarity between the collocates. These

onomatopoeia-based collocations bear the lexical function Son '(to) emit characteristic sound'.

4.4 Adverbial phrase (noun + *-n, -ra, -z*) + verb

In Basque collocations only three postpositions occur significantly: the allative (*-ra* 'to-where'), the inesive (*-n*, 'where/when'), and the instrumental (*-z*).

Examples with the inesive postposition:

martxan/abian jarri [move/start-in place] 'to give/take a start' (LF IncepOper$_1$); *auzi(t)an jarri/ipini* [question-in put/place] 'to call (sth.) into question' (LF Labor); *zalantzan jarri/ipini* [doubt-in put/place] 'to doubt (about sth.)' (LF Labor); *aurrean joan* [front-in go] 'to go in front, lead' (LF Loc$_{in}$).

Examples with the allative:

auzitara eraman [court-all-s carry] 'to take (someone) to court' (LF Labor)*; burura eraman [head-all carry] 'to carry out'* (LF Oper); *atzera bota* [back-all throw] 'to throw back, reject' (LF Oper); *aurrera egin* [forward-all do] 'to go forward, advance' (LF Oper).

Examples with the instrumental:

abantailaz jokatu [advantage-ins play] 'to play with advantage'(LF Oper$_1$); *negarrez urtu* [tear-ins/in melt] 'to melt in tears'; *ahalkez urtu* [shame-ins melt] 'to melt (oneself) with shame'; *urguluz hantu* [pride-ins swell] 'to swell with pride.'

The last three examples bear the lexical function Magn or Excess.

5 Conclusions

We have outlined a classification based on the morphosyntactic structure of the collocations. We have distinguished three types: noun-based, adjective-based, and adverb + verb. We find difficulty in specifying what the base is in many examples of the latter type (adverb + verb).

One noteworthy result of the description is that an important number of the morphosyntactic patterns of Basque collocations are different to the Romance languages of the region. The different patterns we have recognized are the following: noun (ergative) + verb (2.1); noun (dative) + verb (2.2); noun + egin verb (2.3.1); <adjective suffix -ko> + noun (2.5); units with a double and (quasi)synonymic collocate (4.2); and onomatopoeic constructions (4.3.).

From the examples shown we can deduce an abundance of the following constructions: noun (absolutive = object) + verb (2.3.1), <noun + adjective> (2.4.), onomatopoeic collocations (4.3) and adverbial phrase + verb (4.4).

The examples show a diversity of lexical funtions (12 Simple Standard LF and 9 Complex Standard LF), being the most abundant $Oper_1$ and Magn.

The abundant examples we have analyzed indicate the combining possibilities and constraints of many common words, and present a broad panorama which we believe may be of use for Basque speakers and linguists.

Bibliography

Abaitua, J. 1988. *Complex Predicates in Basque: From Lexical Forms to Functional Structure*. (PhD dissertation), Manchester: University of Manchester, Institute of Science and Technology (UMIST).

Alberdi, X., Altzibar, X. & Garcia, J. 2010. Calcos fraseológicos en euskera en los medios de comunicación. [International] Congress: *Europhras 2010. Cross-Linguistic and cross-cultural perspectives on Phraseology and Paremiology*. Departamento de Lingüística General y Teoría de la Literatura. Asociación Andaluza de Lingüística General. European Society of Phraseology. (In press).

Alonso Ramos, M. 2001. Constructions à verbe support dans les langues SOV. *Bulletin de la Société de linguistique de Paris*, t. XCVI (2001), fasc. 1, 79-106.

Altzibar, X. 2005. Kolokazioak euskaraz. Zer axola duten kazetaritzan. *Euskarazko Kazetaritzaren I. Kongresua. Kazetaritza euskaraz: oraina eta geroa*. UPV-EHU 2005. http://www.argia.com/kazetaritza2004xabieraltzibar.pdf

Bosque, I. (dir.). 2004. *REDES. Diccionario combinatorio del español contemporáneo*. Madrid: Ediciones SM.

Corpas, G. 1997. *Manual de fraseología española*. Madrid: Gredos, 1997.

Mel'čuk, I. A. 1998. Collocations and Lexical Functions. In A.P. Cowie (ed.), *Phraseology: Theory, Analysis, and Aplications*, Oxford University Press, 23-53.

Active dictionary of the Russian language: theory and practice

Valentina Apresjan

Russian Language Institute
Volkhonka 18/2, 119019 Moscow
valentina.apresjan@gmail.com

Abstract

The paper is devoted to the Active dictionary of Russian (ADR), which is currently being compiled under the guidance of Yury Apresyan at the Sector for Theoretical Semantics at the Russian Language Institute, Moscow. The paper consists of two parts: 1. General principles of ADR as formulated by Yury Apresyan in the Lexicographer/User Guidelines for ADR; 2. Description of color terms in ADR.

Keywords

Active dictionary, semantics, syntax, co-occurrence properties, color terms.

1 General principles of ADR

ADR is an innovative type of a dictionary which aims at combining the latest achievements of linguistic theory with a practical usefulness for a wide range of language learners. While retaining the best in the European tradition of active dictionaries, ADR is an attempt at its radical modernization in adherence to modern lexicographic principles (integral description of language and systematic lexicographic treatment of kindred linguistic phenomena), with the use of contemporary lexicographic technologies (language corpora and linguistic experiment) as well as the latest theoretical achievements in semantics, syntax, co-occurrence properties and lexicalized prosody. ADR's intended lexical coverage is about 10000 lexical items.

1.1 Semantics in ADR

Semantic innovations in ADR concern (a) semantic definitions; (b) semantic rules.

(a) As for the definitions, they have to meet five major requirements: they have to be systemic, comprehensive, non-tautological, explanatory of co-occurrence properties of lexical items (theoretical requirements) as well as intelligible to a non-professional user (practical requirement).

(b) Under certain contextual conditions, the prototypical meaning of a lexeme as reflected in its definition can undergo changes. Regular semantic modifications, which always occur in the same, strictly verifiable, contextual conditions and can, therefore, be accounted for by semantic rules, are termed different *usages* of a lexeme. These semantic modifications and rules (contexts and the corresponding semantic shifts) are described after the MEANING, in the COMMENTARY part of the dictionary entry.

1.2 Syntax in ADR

(a) Three-level theory of government. The main instrument of describing governing properties of predicates in ADR is Government Pattern (GP), a lexicographic construct, introduced in the Meaning-Text theory of I. Mel'čuk. The starting point for constructing a predicate's GP is its analytic definition, which allows one to establish the number of its semantic valencies. This number equals the number of variables A1, A2, ...,An, which are used in the analytic definition, which, in turn, is determined by the number of obligatory participants in the situation described by that predicate. Thus, in a situation described by the verb *pribivat' / pribit'* 'to nail / to nail down,' there are five obligatory participants: 1) the one who nails (A1, Agent), 2) the object which is being nailed (A2, Patient), 3) the object to which something is being nailed (A3, Patient / Place), 4) the object with the help of which something is being nailed (A4, Instrument), 5) the object which is used as means of fastening (A5, Means). Government is only postulated for cases when semantic valencies of a predicate are filled by words or groups of words that are also its syntactic dependents.

(b) Non-valenced syntactic properties of lexemes. Apart from GP, ADR records a number of non-valenced syntactic properties of lexemes, which trigger their ability or inability to occur in constructions of the so-called "small syntax." A group of Russian existential verbs can provide an example: *byvat'* 'to be,' *byt'* 'to be,' *vodit'sja* 'to be encountered,' *vozniknut'* 'to occur,' *vyjti* 'to happen,' *dut'* 'to blow,' *zavestis'* 'to appear, make one's home,' *imet'sja* 'to exist,' *najtis'* 'to be found,' *proizojti* 'to happen,' *slučit'sja* 'to take place,', *strjastis'* 'to befall,' *suščestvovat'* 'to exist,' etc. The meaning of such verbs is, as a rule, entirely included in the meaning of the subject, or is its pragmatic implicature, thus rendering their own semantic contribution to the meaning of a sentence negligible. The subject thus becomes the rheme of the sentence, as it contains the maximum new information and, as a result, the order of the subject and the predicate is inverted. Instead of the neutral Russian SV order it becomes VS.

1.3 Co-occurrence properties in ADR

ADR is primarily concerned with lexicalized co-occurrence properties, which in MTT are accounted for by the theory of lexical functions (LFs) of I. Melčuk and A. Žolkovsky. ADR uses a modified theory of LFs according to which the choice of a specific word L to fill out a given LF is not arbitrary, but motivated (although not one hundred percent motivated) by a shared semantic component in the lexical meanings of L and X. However, the complicated formal apparatus of LFs remains "behind the scenes"; it is used for a targeted collection of material and its systemic representation in ADR. LFs "surface" in the short semantic characteristics that accompany clusters of phrases in the CO-OCCURRENCE part of the entry. For example, in the entry *travma* 'trauma' (by M. Glovinskaja), MAGN and AntiMAGN would be referenced as 'degree' (*tjaželaja* VS. *ljogkaja travma* 'heavy VS. light

injury'); CAUS as 'causing trauma' (*pričinjat', nanosit' travmu* 'to incur an injury'); LIQU as 'liquidating trauma' *(lečit', zalečit' travmu* 'to treat/to heal an injury'); cf. (Mel'čuk & Polguère 2007), where LFs are also presented informally.

1.4 Lexicalized prosody in ADR

Prosody comprises a wide range of phenomena, out of which ADR is concerned primarily with those that pertain to phrasal stress as the most frequently lexicalized and therefore the most lexicographically interesting prosodic pattern.

Lexicalized phrasal stress is always in some way tied up to the communicative structure of the sentence. There are two groups of prosodic phenomena that are reflected in ADR: a) prosodic syntagmatics; b) prosodic pardigmatics.

a) Prosodic syntagmatics. Certain lexemes, while themselves not prosodically marked, do, at the same time, require prosodic accentuation of the words they are syntactically connected with. Thus, the particle *čto kasaetsja* 'as for X,' 'what concerns X' marks the NP to the right of it, on which it is syntactically dependent, as the contrastive rheme of the sentence, and requires its accentuation with a logical stress.

b) Prosodic paradigmatics. Prosodic paradigmatics deals with prosodic accentuation as a marker of differences among various lexemes of the same word or different usages of the same lexeme. Prosodic accentuation tends to mark the following categories of meanings: 1) negation; 2) quantification; 3) modalities of desire, necessity and possibility; 4) evaluation; 5) facts and opinions. The presence of one or more of those meanings in the semantics of a lexical item allows one to form expectations concerning its prosodic properties.

This phenomenon can be illustrated with two different lexemes of the word *pozdno* 'late.' *Pozdno 1* 'late 1' means 'at a late time' and can either bear phrasal stress or be prosodically unmarked, whereas *pozdno 3* which means 'too late for doing X' and combines the components of quantification, lost possibility, and evaluation, always bears the main phrasal stress.

2. Color terms in ADR

In this part of the paper, I would like to illustrate the above-formulated principles of ADR with the material of color terms. Color terms form a very tightly-knit semantic and lexicographic class, and they display a number of common properties which a systematic lexicographic description allows to reveal. Color terms have been extensively researched in linguistics, and to an extent, semantic principles of their description in ADR are based on Anna Wierbicka's classical treatise (Wierzbicka 1990:99-150). Many important guidelines to the MTT treatment of color terms were also explicitly or implicitly formulated by I. Mel'čuk and Yu. Apresyan in the lexicographic entry of the word *cvet* 'color' in the Explanatory and Combinatorial Dictionary (Mel'čuk & Zholkovsky 1984:931-940).

2.1 Semantic properties of color terms

Within the semantic class of color terms, achromatic colors (*belyj* 'white' and *černyj* 'black') form an important, even though small, subclass. Due to their semantics in their first meaning, these two terms develop a whole range of meanings which are absent in other color terms.

2.1.1 Semantics of 'white' and 'black'

Overall, the polysemies of 'white' and 'black' follow two general patterns: gradual semantic "voiding" typical of peripheral meanings, on the one hand, and development of strongly idiomatic peripheral meanings, on the other.

In their primary meaning, *belyj* 'white' and *černyj* 'black' are defined as 'the color of milk' and 'the color of soot,' respectively. Because of 'white' being the prototypical "light" color and 'black' being the prototypical "dark" color, they develop a number of idiomatic meanings where they are used to refer to light and dark, rather than white and black objects, respectively. Those meanings range from those closest to the prototype to much more semantically diluted and void ones.

The first group of meanings can be illustrated with such examples as *belye ruki* 'white hands' or *černye glaza* 'black eyes,' where the color terms mean 'very light, close in color to white' and 'very dark, close in color to black,' respectively. The second group of meanings can be illustrated with such examples as *belye figury v šahmatax* 'white pieces in chess' vs. *černye figury v šahmatax* 'black pieces in chess' or *belyj hleb* 'white/wheat bread' vs. *černyj hleb* 'black/rye bread,' where the meaning is defined relatively to its opposite in the implied dichotomy of 'non-dark' vs. 'non-light.' The actual color of chess pieces, bread, and many other objects described with the term 'white' can vary anywhere in the light parts of the color spectrum, in the same way as the corresponding objects described as 'black' can be color-wise anywhere in dark part of the spectrum.

One of these semantically "diluted" meanings, the terminological 'white' and 'black,' used for example, in biological, geological and other nomenclature, is inherently *comparative* in nature: 'white' is used to refer to objects that are lighter than other objects of the same type, whereas 'black' is used to refer to the objects that are respectively darker; cf. *belaja fasol'* 'white beans' (white, yellow, pinkish, brownish) vs. *černaya fasol'* 'black beans' (black, dark-brown, dark-red), *belyj čaj* 'white tea' (whitish, greenish, brownish) vs. *černyj čaj* 'black tea' (gray, brown), *belyj kvarc* 'white quartz' (greyish) vs. *černyj kvarc* 'black quartz' (dark brown), etc. Thus, one of the strategies in the semantic development of achromatic color terms is gradual semantic voiding, which takes place in two stages:

1) first, they become general designators of light and dark colors, respectively (as in 'white skin,' 'black eyes,' etc.);

2) further on, they become designators of non-dark and non-light colors, respectively, often in dichotomic oppositions where there are no other color options apart from those two (such as white-black chess pieces, white-black gold, white-black bread, etc.).

This semantic voiding and "abstractization" is paralleled by the development of negation component in the peripheral meanings of achromatic color terms. It is especially strongly

manifested in the polysemy of 'white,' as the prototypically least defined of colors (practically, the absence of coloration); cf. the explicitly negative meaning of the lexeme 'white 4', as in *belyj holst* 'white canvas':

(1) *Belyj A1* 'white A1' (*belaja bumaga* 'white paper,' *belyj holst* 'white canvas') = 'Object A1, on which people usually write texts or draw images, uncovered by text or images and therefore close in color to white.'

Negation is also apparent in other meanings of 'white,' such as in *Ego volosy stali uže soveršenno belymi* 'His hair has already gone completely white' (defined in (2)) or *Ego lico stalo soveršenno belym* 'His face has gone completely white' (defined in (3)):

(2) 'Having lost natural pigmentation and having acquired a color close to white due to an old age or for another reason'
(3) 'Having lost natural coloration and having acquired a color close to white due to a strong and unpleasant emotion or sensation A2' [usually about a person's face or hands]

The second strategy, typical for the polysemy of 'white' and 'black' in Russian, is opposite to semantic voiding, and consists in the development of strongly idiomatic meanings containing positive and negative evaluations, respectively. If semantic voiding is triggered by the "real world" denotation of these color terms, their figurative meanings stem from their cultural connotations.

In Russian, as in other European languages, 'white' bears the positive connotations of light, goodness, purity, innocence, whereas 'black' carries the connotations of badness and dishonesty. These connotations give rise to multiple figurative meanings; to name a few examples of parallelism between 'white' and 'black,' let us consider the lexemes represented in the following examples: *Nel'zja delit' mir na černoe i beloe* 'One cannot divide the world into black and white'; *belaja magija* 'white magic' vs. *černaja magija* 'black magic,' *belaja zavist'* 'white <non-malicious> envy' vs. *černaja zavist'* 'black <malicious> envy.' In these essentially LF meanings of Bon and AntiBon, respectively, 'white' and 'black' are extremely lexically restricted. The meanings of corresponding 'white' and 'black' lexemes are presented in (4a) and (4b):

(4a) 'Morally good, devoid of any bad elements'
(4b) 'Morally bad, devoid of any good elements'

While the definitions might at first glance seem redundant, I consider the components 'devoid of any bad elements' and 'devoid of any good elements' to be of paramount importance: they account for the "extreme," "polar" associations of the white vs. black moral division. While something can be quite, but not entirely, good, or quite, but not entirely, bad, moral 'whiteness' or 'blackness' are decidedly non-gradual.

Another important figurative meaning of 'white' and 'black' based on their 'good' vs. 'bad' connotation is represented in phrases like *belaja zarplata* 'white <legal> salary,' vs. *černaya zarplata* 'black <illegal> salary,' *belyj nal* 'white <legal> cash' vs. *černyj nal* 'black <illegal> cash.' This meaning of corresponding 'white' and 'black' lexemes is explicated in (5a) and (5b):

(5a) 'Accomplished or functioning without breaking financial, primarily tax, laws and therefore not concealed'
(5b) 'Accomplished or functioning by breaking financial, primarily tax, laws and therefore concealed'

Apart from the 'good' vs. 'bad' connotation, this meaning stems from another connotation these two color terms possess, namely, that of light vs. darkness. 'White' in the meaning of 'legal' describes easily observable, unconcealed activities, that are exposed to the "light" of everybody's scrutiny, whereas 'black' in the meaning of 'illegal' describes hidden, unobservable activities that are deliberately left "in the dark."

2.1.2 Semantics of other color terms in ADR

There are certain semantic tendencies that are characteristic of all color terms, and one of them consists in developing a meaning of sudden (and usually abnormal) facial color change, which is due to an emotional or physical factor. Adjectival lexemes with this meaning always produce a derivative verb with the meaning of facial color change. This meaning is found in all basic color term adjectives, such as *belyj* 'white,' *černyj* 'black,' *krasnyj* 'red,' *rozovyj* 'pink,' *sinij* 'dark blue,' *zelenyj* 'green,' *seryj* 'gray,' *želtyj* 'yellow,' *bagrovyj* 'purple' and some others. All the color terms that are used to describe facial color change are in some way associated with the increased (pink, red, black) or dicreased (yellow, blue, white, green) blood flow. Most colors tend to describe both emotion-associated changes and changes associated with the influence of physical factors, such as heat, cold, lack of oxygen, disease, etc. This is probably due to the fact that emotions themselves trigger physiological changes, e.g., fear acts as cold, anger acts as heat, etc., which is reflected in the metaphorical conceptualizations of these emotional states (consider classical work by (Lakoff and Johnson 1980) and a later treatise (Kövecses 2000)). Some colors, like *sinij* 'dark blue' tend to refer primarily to changes induced by physical factors. Interestingly, certain basic color terms are absolutely unable to develop this type of meaning; notably, it is impossible for *koričnevyj* 'brown,' which can be used to describe a more permanent skin coloring. It is mostly impossible for the majority of "non-basic" color terms, derived from the names of corresponding objects, such as *sirenevyj* 'lilac,' *limonnyj* 'lemon' and many others. The following generalized semantic form is proposed for this type of meaning, which can be ascribed to most of the basic color terms:

(6) *A1 is of the color X because of A2* 'A part of the human body A1 or a part of the body of the human A1 lost its natural color and acquired color X because of an emotional or physical state A2' [usually about the face or hands]

This is essentially an LF meaning, as it refers to the symptom of the corresponding emotion or physical state. Syntactically, this adjectival lexeme, as well as its derivative verb possess the valency of cause: *belyj ot užasa* 'white of terror,' *belet' ot užasa* 'to go white with terror, etc.'

2.2 Derivational properties of color terms in ADR

As mentioned before, most adjectival color terms derive a verb with the meaning of 'to acquire color X.' This verb has, in its turn, multiple meanings. The typical polysemy of such verbs in Russian is as follows (it is by no means exhaustive, since many verbs have individual additional meanings):

X-t' 'to acquire the color X, to X-en':
1. 'to acquire the color X because of factor A2' [about objects]: *Jabloki krasneli na solnce* 'Apples were reddening in the sun'
2. 'to acquire facial or bodily color X because of a physical or emotional factor A2' [about people]: *Ego lico pokrasnelo ot styda <ot žary>* 'His face reddened because of shame <heat>'
3. 'to appear X-colored to an observer': *V pole krasneli maki* 'literally: Poppies were reddening in the field; Poppies were showing red in the field'

In their second meaning, that of emotion-induced facial color change, these de-adjectival verbs have been extensively lexicographed in the following sources: Iordanskaja's lexicographic entries of emotion words (in Melčuk & Žolkovsky 1984), (Iordanskaja & Melčuk 2007:338-340) (as Sympt Excess^{color} and Sympt STOP^{color} LFs) , as well as in (Iordanskaja & Paperno 1996).

The third, stative, meaning of these verbs is specific to the Russian language. In English, the verbs with the primary meaning of color change, such as *to whiten, to blacken, to redden* can only be used to describe processes. This specific meaning has been previously referenced in linguistic literature (Yu. Apresyan 1995:643, Yu. Apresyan 2000:229). Yury Apresyan formulated certain important properties of verbs denoting perception of colored objects, namely:

1) they all contain a reference to an outside observer, different from the speaker; thus, phrases like ?? *Moe telo belelo v temnote* 'My body was showing white in the dark,' are impossible or extremely awkward, while phrases like *Ee telo belelo v temnote* are perfectly possible 'Her body was showing white in the dark.'

2) the visible object cannot be human (it can be part of a human body); thus, phrases like **Marija belela v temnote* *'Mary was gleaming white in the dark' are impossible.

ADR approach has allowed us to make further adjustments to the linguistic treatment of this group of verbs. This meaning is present in a large group of color change verbs including but not limited to *belet'* 'to turn/appear white,' *černet'* 'to turn/appear black,' *krasnet'* 'to turn/appear red,' *želtet'* 'to turn/appear yellow,' *zelenet'* 'to turn/appear green,' *sinet'* 'to turn/appear dark blue,' *golubet'* 'to turn/appear light blue,' *rozovet'* 'to turn/appear pink,' *bagrovet'* 'to turn/appear purple,' *pestret'* 'to appear colorful, to make splashes of color' and some others. In their stative meaning, most of these verbs have reflexive counterparts with the same meaning; cf. *belet'sja* 'to appear white,' *černet'sja* 'to appear black,' etc.

Below are some examples:

(7) *Na gorizonte beleli bašni zamka*
On horizon whitened towers castle-GEN
'The castle towers gleamed white on the horizon'
(8) *V temnote <skvoz' tuman> beleli stvoly berjoz*
In darkness whitened trunks birch trees-GEN
'Birch tree trunks shone white in the darkness <through the fog>'
(9) *V trave beleli romaški*
In grass whitened daisies
'Daisies were shining white in the grass'

(10) *Na stole v solonke belela sol'*
On table in saltbox whitened salt
'The salt was gleaming white in the saltbox'

The definition proposed for this group of verbs in their stative meaning by Yu. Apresyan in Lexicographer/User Guidelines for ADR is as follows:

(11) *A1 X-eet* 'A1 is X-ing' 'A1 appears X' 'Object A1 of color X is visible to an observer from some distance'

However, the examples above and general corpus study of this class of verbs suggest that the meaning might be formulated more precisely. It seems that the distance is not a necessary pre-requisite for using a verb of this group; in fact, sentences like *Na stole v solonke belela sol'* 'The salt was gleaming white on the table' or *Na tarelke želteli slivy* 'Plums were glistening yellow on the plate' do not point to an observer who is distanced from the object. There are, however, certain situational requirements which have to be fulfilled to justify the use of a verb of this type.

First of all, it is, as formulated in Yu. Apresyan's definition, the presence of an observer. Secondly, as pointed out by Yury Apresyan, the object has to be non-human. This point can be further elaborated upon, as the object cannot be an animal, either; phrases like **Sobaka belela v temnote* 'The dog was shining white in the dark' are equally impossible. At the same time, phrases describing insects, such as *V trave beleli svetljački* 'Fireflies were glowing white in the grass' are considerably better. One might be tempted to say that the ability to co-occur with such verbs is triggered by the degree of animateness – the less animate, the better co-occurrence; however, sentences referring to cars or airplanes appear pragmatically awkward: ?*Na doroge krasneli mašiny* ?'Cars were showing red on the road' or ?*V nebe černeli samolety* ?'Airplanes were showing black in the sky.'

It seems that the second important requirement is that the object be *stationary*; therefore, the best candidates for these contexts are the prototypically immobile objects such as buildings or plants. People, animals, means of transportation are not typically perceived as immobile elements of the landscape; on the other hand, insects sometimes can be, as well as body parts or even whole bodies.

Thirdly, the verbs of this group are stylistically marked, namely, they can only occur in texts of narrative register.

Fourth, the object ought to be visible to an observer *due to its color*. Let us explain this last statement. These verbs, though essentially of an LF type, do add some semantic flavor to the general existential/locative proposition. If the speaker, instead of saying 'There was salt in the saltbox' or 'The salt was in the saltbox' chooses to say 'The salt was gleaming white in the saltbox,' (s)he adds certain information to the utterance.

Note that all the examples include a reference to the place where the object is located. Though this reference is a requirement for existential and locative sentences, in the case of these verbs the requirement is more specific. Consider the following pragmatically awkward phrases:

(12) ??*V cvetochnom magazine želteli hrisantemy*
In flower shop yellowed chrysantemums
??'Chrysanthemums were shining yellow at the florist's'
(13) ??*V lesu zeleneli derev'ja*
In wood greened trees
??'Trees were gleaming green in the wood?

What makes these sentences pragmatically inappropriate? It seems that the requirement for using verbs of this group is that they should refer to objects which are visible against some contrastive background (as 'in the darkness') or through an obstacle (as in 'through the fog'), and visible due to their color. Sometimes, the objects might be at a distance from the observer (e.g., 'Towers were shining white on the horizon'); sometimes they are close (e.g., 'Birches were shining white in the dark'); what matters, is the background or obstacle against or through which they are seen due to their color. If this condition is absent, as in (12), where 'the florist shop' is not a *background*, or (13), where the green 'wood' is not a *contrastive* background for trees, their use becomes impossible. Thus, the corrected definition for this group of verbs as proposed for ADR sounds as follows:

(14) *A1 X-eet v A2* 'A1 is X-ing in A2' 'A non-moving object A1 of color X is visible to an observer against a contrastive background or through an obstacle A2'

This definition includes an extra valency – namely, that of a background (obstacle), which is syntactically expressed in a variety of ways. The more traditional locative expressions include the prepositional noun phrase with 'in' + PREP, as in *v temnote* 'in the dark'; there are also ones that point to an obstacle, as in *skvoz' tuman* 'through the fog,' *za elkami* 'behind the fir trees.'

2.3 Syntactic properties of color terms in ADR

Color terms display a number of uniform syntactic properties that often go unnoticed by traditional dictionaries. Due to space limitations, only one of those properties, namely, the construction with the instrumental case in the meaning of 'part' is going to be described here. It possesses, however, a certain degree of universality in relation to Russian color terms and should therefore be recorded in their ADR entries. At least three groups of color terms, namely, adjectives of color, verbs of color change in their second meaning (change of facial color due to an emotional or physical factor) and verbs of color change in their stative meaning (being observable) can occur in the construction with instrumental in the meaning of 'part':

(15) *Ona byla polna i bela licom*
She was plump and white face-INSTR
'She was plump and white in the face'
(16) *Ona pobelela licom*
She whitened face-INSTR
'Her face went white'
(17) *Cerkvi beleli kolokol'njami*
churches whitened bell-towers-INSTR
'Churches were gleaming with their white bell-towers'

Similar examples can be given for other color adjectives and verbs; cf. *krasnyj licom* 'red in face,' *černet' licom* 'go black in the face,' *zelenejuščie lužajkami londonskie prigorody* 'London suburbs gleaming green with lawns,' etc. This construction illustrates the phenomenon of splitting a semantic valency, in this case, the subject valency. While many verbs exhibit splitting, it has not been previously lexicographed for color verbs. As for the adjectives, the ability to govern an NP in the instrumental case is relatively unusual, and the fact that color adjectives possess it, is revealing. This syntactic ability of color terms to govern an NP in INSTR with 'part' interpretation might be rooted in the semantics of color. This construction describes the part of an object which is most noticeable due to its color. Color is, as noted in the above-mentioned description of the word *tsvet* 'color' in the Explanatory Combinatorial dictionary, a distinct visual characteristic. It is, one might add, one of the most striking and noticeable visual characteristics. And the fact that visual perception is the most important type of perception for humans might account for this special syntactic ability that color terms possess. After all, Russian does not have expressions like **nežnyj golosom* 'tender in voice' or **pušistyj volosami* 'fluffy in hair,' with auditory or tactile lexis. There are other, semantically similar classes of verbs, which possess the same syntactic property. These are verbs of 'light', such as *blestet'*, *sverkat'* 'to gleam, to glitter,' *sijat'* 'to radiate light,' *goret'* 'to burn.' Among their many meanings, there is a meaning in which they denote 'being visible due to reflecting bright light,' as in *Cerkvi blesteli kupolami* 'Churches were glittering with their cupolas,' where the same syntactic phenomenon is manifested, and for the same semantic reasons: the part of object which is most noticeable due to being the brightest, gets its own syntactic expression by splitting the semantic valency of the subject.

2.4 Co-occurrence properties of color terms in ADR

While color terms possess many noteworthy co-occurrence properties, perhaps their most interesting combinatorial characteristic stems from their obligatory correlation with a certain prototype. Each color correlates to an object in the real world which possesses this color in its purest form. Some of these objects are cross-culturally universal, which makes defining color terms possible (e.g., white is the color of milk, black is the color of soot, red is the color of blood, etc.), while others are culturally specific. In Russian, for example, the prototypical white objects, besides milk, are snow, cream, marble, pearl, sugar, and some others, and the word *belyj* 'white' co-occurs with the names of these objects in comparative constructions: *belyj kak sneg* 'white as snow,' *belee snega* 'whiter than snow,' etc. This property of color terms has repercussions on other linguistic levels as well: thus, color terms have adjectival synonyms derived from nouns denoting these prototypical objects, cf. *moločnyj* 'milky,' *slivočnyj* 'creamy,' *snežnyj* 'snowy,' *mramornyj* 'the color of marble,' *saharnyj* 'the color of sugar,' *alebastrovyj* 'alabaster,' etc. They also form compound color adjectives incorporating the name of the object and the correlating color: *moločno-belyj* 'milky-white,' *slivočno-belyj* 'creamy-white,' *snežno-belyj* 'snowy-white,' *perlamutrovo-belyj* 'pearly-white,' etc. All these properties apply to the majority of color terms, cf. the triads *krasnyj kak ogon'* 'red as fire', *ognennyj* 'fiery', *ognenno-krasnyj* 'fiery-red', *černyj kak ugol'* 'black as coal', *ugol'nyj* 'coaly', *ugol'no-černyj* 'coaly-black', *goluboj kak nebo* 'blue as the sky', *nebesnyj* 'the color of the sky', *nebesno-goluboj* 'sky-blue,' etc.

2.5 Polysemy of color terms in ADR

The last illustration of how color terms are treated in ADR is a synoptic outline of all the meanings of the word *belyj* 'white,' which, although not entirely applicable to other color terms or even to its closest counterpart and antonym *černyj* 'black,' still captures the general logic of the development of "color" semantics. The outline is presented below. In this outline, we can see the gradual dissolution of the prototypical meaning, via the meaning 'light color' (second block) and the meaning 'devoid of natural coloration and therefore light' (third and fourth blocks), to the meaning 'non-dark' (sixth and seventh blocks) and, finally, to the meaning 'good' (eight block). This development mirrors the increasing semantic bipolarity and the corresponding loss of gradability: while adjectives in the first three blocks are gradable *(Ee ruki belee moih* 'Her hands are whiter than mine'), starting with the fourth block they are decidedly non-gradable (**Èti griby belee teh* *'These mushrooms are whiter than those').

belyj 1 'the color of milk': *beloje plat'je* 'white dress'.
belyj 2.1 'of a color close to white': *belye ruki* 'white hands'.
belyj 2.2 'belonging to the race of people with a relatively light skin [often substantivized]': *belyj čelovek* 'white person'.
belyj 3.1 'having lost natural coloration and having acquired a color close to white due to a strong and unpleasant emotion or sensation': *belyj ot straha* 'white of fear'.
belyj 3.2 'having lost natural pigmentation and having acquired a color close to white due to an old age or for another reason': *belye volosy* 'white <gray> hair'.
belyj 4 'uncovered by text or images and therefore close in color to white': *belyj holst* 'white canvas'.
belyj 5 'characterized by natural lighting': *belyj den'* 'white <light> day'.
belyj 6 'starting the game of chess or checkers and having a lighter color than the pieces that do not start the game': *belyj ferz'* 'white queen'.
belyj 7.1 'having a lighter color that other objects of the same class or a white color': *beloje vino* 'white wine'; *belyj medved'* 'white <polar> bear'.
belyj 7.2 'white <porcini> mushroom' [substantivized]: *V ijune pojavilis' pervye belye* 'First white mushrooms <porcini> appeared in June'.
belyj 8.1 'morally good, devoid of any bad elements' ': *belaja magija* 'white magic'.
belyj 8.2 'accomplished or functioning without breaking financial, primarily tax, laws and therefore not concealed': *belaja zarplata* 'white <legal> salary'.
belyj 9 'related to the Russian counter-revolutionary movement whose aim was the restoration of monarchy': *Belaja Armija* 'White <counterrevolutionary> Army'.

Acknowledgements

This paper was partly supported by the following grants: grant of the Program for Fundamental Research OIFN of the Russian Academy of Sciences "Genesis and interaction of th social, the cultural, and th lingustic," grant of the Russian State Humanities Fund № 10-04-00273a "Preparation of the first volume of the Active Dictionary of Russian," grant for Scientific Schools NSH-4019.2010.6 for supporting research conducted by the leading scientific schools of the Russian Federation.

Bibliography

Apresyan, Yuri. *Selected works. v. 2. Integral description of language and systematic lexicography.* Moscow, 1995.

Apresyan, Yuri. *Systematic lexicography*. Translated by Kevin Windle. Oxford University Press, 2000.

Apresyan, Yuri. *Lexicographer/User Guidelines for the Active Dictionary of Russian Language* (in print).

Iordanskaja, Lidija & Igor Mel'čuk. Smysl i sočetaemost' v slovare. *Jazyki slavjanskih kul'tur*. Moscow, 2007.

Iordanskaja, Lidija & Slava Paperno. A Russian-English Collocational Dictionary of the Human Body, ed. by Richard L. Leed, Slavica Publishers, 1996.

Kövecses, Zoltan. *Metaphor and Emotion*. Cambridge University Press, 2000.

Lakoff, George & Mark Johnson. *Metaphors we live by*. Chicago & London, The University of Chicago Press, 1980.

Mel'čuk, Igor & Alain Polguère. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20.000 dérivations sémantiques et collocations du français.* Bruxelles: De Boeck, 2007.

Mel'čuk, Igor & Alexander Žolkovsky. *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-syntactic Studies of Russian Vocabulary.* Vienna: Wiener Slawistischer Almanach.

Wierzbicka, Anna. The Meaning of Colour Terms: Semantics, Culture and Cognition". In *Cognitive Linguistics,* 1, 1. 1990. 99-150.

Semantic Analysis Based on Linguistic and Ontological Resources

Igor Boguslavsky

Artificial Intelligence Department – Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain /
Computationl Linguistics Lab – IITP RAS
19, Bolshoj Karetnyj pereulok, GSP-4, 129447 Moscow

igor.m.boguslavsky@gmail.com

Abstract

We describe a recently launched project whose objective is to develop an advanced converter of natural language text to semantic structures. The project basically consists in enriching the ETAP-3 linguistic processor, developed by the Institute for Information Transmission Problems, Russian Academy of Sciences, with a new module – that of semantic analysis. An important feature of this module is that it will use not only linguistic knowledge incorporated in the grammar and the combinatorial dictionary, but also extralinguistic knowledge stored in the ontology and contextual information accumulated in the fact repository. We developed a small ontology that serves as the semantic metalanguage for Semantic Structures. Several examples are given that show how the ontology is used and how the meaning is represented.

Keywords

Semantic analysis, ontology, Semantic web, semantic rules, semantic structure

1 Introduction

Modern search engines such as Google, Yahoo, or Russian Yandex, have long come into our everyday life and we hardly imagine how we could do without them. Nevertheless, however useful these applications may be, they are rightfully reproached for "not understanding" the texts they are dealing with. They find far too many texts, while the overwhelmingly most part of them has nothing to do with what the user is asking about. On the other hand, if a text conveys the relevant meaning but it is expressed by words different from the ones used in the user's query, this text will hardly be found at all. For many NL applications, first of all, for Information Retrieval and Extraction as well as for Question Answering, it is essential that

they should be able to discover semantic similarity between the texts if they express the meaning in different ways.

As is known, most of the web content is created to be read by humans and not for meaningful operations on this content by machines. This is what brought to life the Semantic Web initiative which aims at making the web content more understandable for computers (Berners Lee et al., 2001). This requires, in particular, the elaboration of methods of representing semantic information contained in the texts and creating efficient technologies of handling it. However, there is a serious problem here. Annotating texts with semantic markup and developing resources is costly, and hardly anybody would be willing to work on that on a large scale unless there are valuable applications that will use these resources. But at the same time, applications are not likely to be developed before there exist semantic markup and resources. This is one of the reasons why Semantic Web is slow to realize its potential.

A natural way out of this impasse is automatic or semi-automatic semantic processing of NL texts. Advanced NLP systems will play a twofold role here. On the one hand, they could produce semantic annotation of texts on the massive scale, and on the other hand, they will benefit from the created resources themselves (Pall, 2006). To some extent, the situation is similar to the creation of treebanks. To build a large treebank, one needs an effective parser, but once a treebank exists, it can be used for creation of new parsers as well as for testing and evaluation of the existing ones.

Many NL applications need a much deeper semantic analysis of the text than is used today. In section 2 we will show with a simple example what perspectives deep semantic and encyclopaedic information opens for Information Retrieval and Question Answering. To make use of this information one has to rely on both linguistic and extralinguistic resources. In subsequent sections we will describe a system under construction at the Institute for Information Transmission Problems of the Russian Academy of Sciences and at the Technical University of Madrid which aims at combining both types of resources in order to produce semantic structures to be used in different applications. In section 3 we will briefly present the ETAP linguistic processor, and in section 4 – the SemAnOn project, which is responsible for developing the semantic analysis module for ETAP. Section 5 will highlight the domain ontology developed for the project. In section 6 we show that even simple class/subclass and part/whole relations can be effectively used for syntactic disambiguation and co-reference resolution. Section 7 will illustrate with a series of examples the kind of semantic structures we are striving for. In section 8 we list some directions of future research.

2 Information Extraction and Question Answering in Need of Semantic Analysis

Suppose we wish to find information about

(1) *the losses of warships during World War II.*

If we want to solve the task with Google, we have two main options. First, we could try “this exact wording” option. In this case, no match will be found. Then, we could look for the texts that contain all the words of the query, although not necessarily in the same order and maybe in different sentences. In this case, we obtain more than 16.800.000 links. Most of the texts

found are noise, irrelevant for our quest. On the other hand, many of the documents which are relevant will be buried among hundreds of thousands of irrelevant texts, which is tantamount to not being found at all.

One of the ways for increasing the accuracy of information retrieval is taking into account the syntactic structure of both the query and the candidate text. If the query is *Working plan*, the relevant response should not be *We are working on the plan*, because syntactic relations between *working* and *plan* are different in the query and in the response. In our warship example, the syntactic structure of the query manifests the temporal relation between the *losses* and *World War II*. Therefore, we should skip all the texts where these words are not connected in this way.

However, the syntactic structure is by far not sufficient for real understanding of the query. We should also be aware that *the loss (of the ship)* means that the ship ceased to exist ($FinFunc_0$), and for this reason belongs to the same semantic class as the nouns *death, wreck, crash,* etc. These nouns are (quasi)synonyms and could be found in WordNet. But we should also understand that there exist a number of verbs that denote the same situation, such as *to die, to sink, to perish, to wreck,* etc. (Note that the noun-verb type of synonymy is not reflected in WordNet). Moreover, we need not only the words that denote the same or a similar situation. An important semantic relation is entailment. In our case, it is essential to know that the end of existence of a thing can be a result of different events of the LiquFunc0 class, such as *to destroy, to kill, to explode, to sink, to eliminate, to exterminate, to liquidate,* etc. A semantic structure that conveys all this information could look like this.

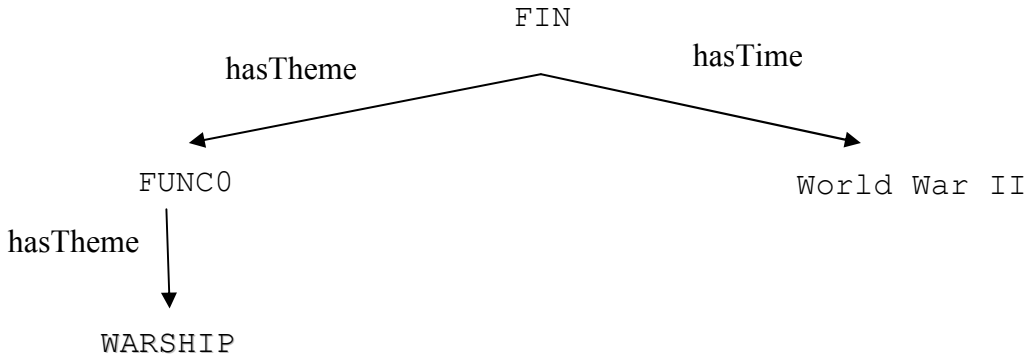


Fig. 1 Semantic structure of query (1): *the losses of warships during World War II.*

Besides semantic information that pertains to linguistic competence and should be incorporated in the linguistic description, there are other important sources of knowledge necessary for understanding the text. This is first of all the encyclopaedic knowledge (contained in an ontology) and an inference mechanism responsible for drawing conclusions from all the information available.

Let us show how the combination of these resources helps ensure a more accurate processing of our example. First, we should enrich the semantic structure in Fig. 1 with encyclopaedic information. Given that World War II began in 1939 and ended in 1945, the representation of the phrase *during World War II* can be made more precise; cf. Fig. 2.

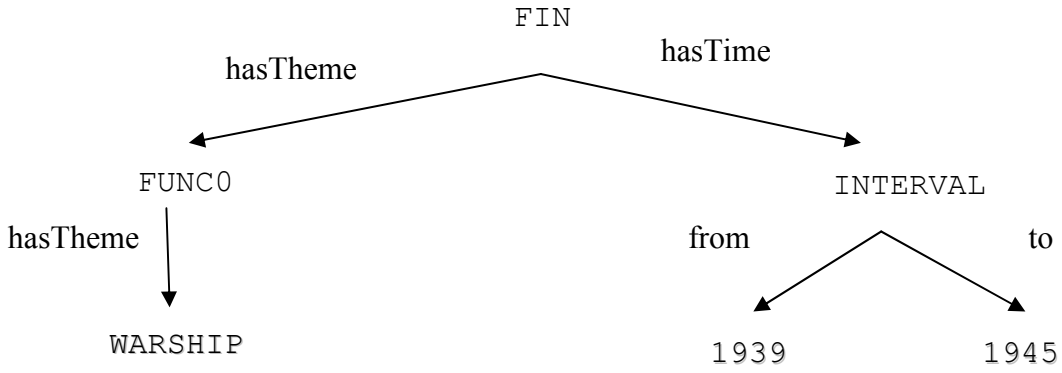


Fig. 2. Semantic structure of query (1) enriched with encyclopaedic information.

Now it is not difficult to see that texts (2) and (3) correspond to query (1), although they are composed of totally different words and are therefore completely beyond the possibilities of word-based retrieval methods.

(2) *On May, 27 of 1941 the Royal Navy destroyed the German battleship "Bismarck".*

To make sure that (2) corresponds to the semantic structure in Fig. 2, one should have access to the following information:

(a) *battleship* is a kind of a warship;

(b) *X destroyed Y* entails that 'Y ceased to exist';

(c) the date *May 27, 1941* belongs to the interval between 1939 and 1945.

This information by far surpasses the data available in NLP systems. Although the data on class-subclass relations, as manifested in (a), can be found in WordNet, (b) requires semantic decomposition rules, and (c) involves reasoning about time.

To come to the conclusion that phrase (3) also corresponds to query (1), one needs to consult an encyclopaedia and find that HMS Hood is a battle cruiser, which is also a member of the warship class.

(3) *the unexpected wreck of Hood at the 8th minute of the battle on May, 21, 1941.*

At the same time, comparing semantic structure of the query with the semantic structures of the texts permits to disqualify the texts which do not match the query. Cf. (4) which, as well as (2) and (3), refers to a wreck of a warship, but in a wrong time span.

(4) *The second film of the series is dedicated to the wreck of the 57-canon motor sailing frigate "Oleg", perished in August 1869.*

Concluding, to implement semantic search, the following resources are necessary:

- a sophisticated linguistic model. It should not only cover domain terms, but also be able, in particular, discover semantic identity in different syntactic contexts (cf. *the wreck of the frigate vs. the frigate whose wreck became known*...). It should be able to handle collocations, modality, and negation; cf. *suffered* [$Oper_1$] *wreck* (relevant for query (1)) vs. *avoided a wreck* (irrelevant)*; didn't suffer wreck* [neg + $Oper_1$] (irrelevant) vs. *didn't avoid a wreck* (relevant);

- various external ontological and encyclopaedic resources and the capacity to integrate them with NLP modules (WordNet, SUMO, FreeBase, DBpedia);

- logical inference engine;

- a large semantic index, which covers the corpus where the search is supposed to be carried out.

3 ETAP-3 Linguistic Processor

The multifunctional ETAP-3 linguistic processor, developed by the Computational Linguistics Laboratory of the Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, (see e.g. Apresjan *et al*. 2003), is the product of decades of research and development in the field of language modelling. At the moment, ETAP-3 consists of a number of options, including

1) a rule-based machine translation system working both ways between Russian and English (plus several prototypes for other languages – French, German, Spanish, Korean and Arabic);

2) a system of synonymous and quasi-synonymous paraphrasing of sentences (Apresjan, Tsinman, 2002);

3) an environment for deep annotation of text corpora, in which SynTagRus, the only corpus of Russian texts tagged morphologically, syntactically (in the dependency tree formalism), and lexically was created (Boguslavsky et al., 2000a), and

4) a Universal Networking Language (UNL) module, responsible for automatic translation of natural language text into a semantic interlingua, UNL, and the other way around (Boguslavsky et al., 2000b).

The ETAP-3 processor is largely based on the general linguistic framework of the Meaning ⇔ Text theory by Mel'čuk. An important complement to this theory was furnished by the theory of systematic lexicography and integrated description of language proposed by Jurij Apresjan (Apresjan, 2000).

For each sentence the processor successively builds several representations: Morphological Structure (MorphS), Syntactic Structure (SyntS), and Normalized Syntactic Structure (NormSyntS). Some options of ETAP-3 make use of a deeper representation. It is a UNL Graph in the UNL option, and Semantic Structure in the semantic analysis option, described below.

One of the major resources used in ETAP-3 is the combinatorial dictionary. It offers ample and diverse data for each lexical entry. In particular, the entry may list the word's syntactic and semantic features, its subcategorization frames, as well as rules (or reference to rules) of a dozen types, which make it possible to describe peculiar behaviour of individual words and exceptions to general rules in a complete and consistent way. Many dictionary entries contain information on lexical functions.

The entry of the combinatorial dictionary has a number of zones, one of which provides the properties of the word that are manifested in the given language, while all the other zones contain information on the match between this word and its equivalent in a particular language. For example, the EN zone in the Russian combinatorial dictionary entry contains information on the translational equivalents of the respective Russian word into English. One field (TRANS) gives the default single-word translation (or several such translations) of this word in English. Other fields contain less trivial translation rules, or references to such rules.

A newly introduced ONTO zone offers information underlying the match between the Russian word and its counterparts in the ontology.

4 SemAnOn project

A new project underway at IITP is aiming at moving ETAP further towards semantics. We add a new module – that of SEMantic ANalysis based on ONtology (SemAnOn). Its task is to transform NormSyntSs into Semantic Structures (SemS).

Our analyser is rule-based, which may seem bizarre nowadays when most of the analysers, both shallow and deep, are statistical. Cf., for example, interesting results obtained in supervised and unsupervised semantic parsing in Ge and Mooney, 2005; Poon and Domingos, 2009; Clark et al. 2010. A combination of machine learning and rule–based approaches is used for semantic processing in Moldovan et al., 2010. Our choice of the strategy is based on two considerations. First, there exist no corpora annotated with the kind of structure we are interested in. Once we construct our analyser, it will open the possibility to develop such a corpus, which could then be used for refining and evaluating the analyser, as well as for developing other semantic parsers. The second, even more important, reason for our non-statistical approach is our firm belief that the modelling of real understanding of texts requires knowledge-intensive methods.

Our approach to semantic analysis is closely related to the OntoSem approach (Nirenburg, Raskin, 2004), although linguistic frameworks adopted in these projects are largely different. Semantic analysis will be done in two steps. First, Basic Semantic Structures are produced, which present literal meaning of the sentence to the extent it can be extracted from the sentence itself. Then, they are transformed into Extended Semantic Structures, that are enriched with ontological and contextual information available.

We are going to use two new types of resources: an ontology and a fact repository The ontology is a collection of concepts connected with relations and provided with attributes and rules. The fact repository accumulates semantic structures which store data about concrete situations.

These resources are to be used for two purposes. On the one hand, they are the basis of semantic analysis. As the matter of fact, ontology provides the metalanguage of semantic structures, so that semantic structures interpret natural language sentences in terms of the ontology. On the other hand, they can be used for disambiguation.

In this paper, we will not give a systematic description of the SemAnOn project. Instead, we will give several examples that illustrate some of its aspects. More details on the project can be found in Boguslavsky et al. 2010.

In the next section, we will demonstrate the small domain ontology we built for the project.

5 Football ontology

The ontology we are working with focuses in the first place on football. It contains information on teams, players, football field, sport events, and their properties. However, we want it to be extendable to other sports as well. That is why some classes are more general than would be needed for football alone. For example, instead of having one class `FootballPlayer`, the ontology has a more general class `Sportsman`, of which `FootballPlayer` is a subclass. An equivalence restriction states that `FootballPlayer` is a `Sportsman` whose `SportType` is `football`. In this way, sportsmen doing different types of sports can be treated by the ontology in a uniform way.

The football ontology is written in SWRL (Semantic Web Rule Language), which is OWL (Ontology Web Language) augmented with rules (Horrocks et al. 2004). In compiling it, we used some existing ontologies dealing with football (e.g. http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml). As usual, properties of classes are inherited by the subclasses. For example, the `Match` class is a subclass of `SportEvent`, which in its turn is a subclass of `Event`. `Match` inherits from `Event` the properties of having definite `Time` and `Place`. From `SportEvent` it inherits the fact that its participants should be `SportAgents`. Its own properties are: the number of participants is 2 (as opposed to championships, which have more) and it has a definite sport type (as opposed to Olympics, which involve many sport types). A subclass of `Match` is `Derby`, in which both participants should be from the same city or region. This property is implemented by means of a SWRL rule. Another rule assigned to `Match` states that if its `SportType` is football (or any other team sport), then its participants should be teams and not individual sportsmen, as is the case in tennis or chess. `Sportsman` is a subclass of two classes: `Person`, from which it inherits the property of having a name and a birth date, and `SportAgent`, which includes also `Team` and from which it inherits the property of having a definite sport type and a coach.

6 Ontological information in parsing

As is well-known, an important source for the increase in the parsing quality is restrictions on the semantic class of arguments of predicates. This information exists in the ETAP-3 combinatorial dictionary for a long time. The restrictions are based on the mini-thesaurus that contains several dozen concepts. In the future, we hope to replace it with a full-fledged upper

and middle ontology. However, besides checking semantic agreement between the predicate and its arguments, there are many other tasks that are solved at the stage of parsing and that can be enhanced by using information available in the ontology. In the first place, it is the resolution of syntactic and lexical ambiguity. Let us illustrate it with one example, which manifests syntactic ambiguity:

(5) *Zaly dlja priemov inostrannyx delegatsij, steny kotoryx byli ukrašeny lepninoj* 'halls for the reception of foreign delegations whose walls were covered with stucco moulding'

The relative clause 'whose walls were covered with stucco moulding' can be syntactically linked to any of the three preceding nouns – 'halls', 'reception' и 'delegation'. In order to select among them the one which is most plausible semantically, we have to check which of the phrases – *walls of the halls*, *walls of the reception* or *walls of the delegations* – makes more sense. Obviously, it is the first of them that can be more easily comprehended, although on terms of word order *halls* is further than other candidates from the relative clause. The semantic agreement between 'halls' and 'walls' can be easily discovered by means of the ontology. Indeed, concepts `Hall` and `Wall` are linked in the ontology by a short semantic chain consisting of relations "is a" и "part of": a `Wall` is a typical part of `Premises`, and a `Hall` is a kind of `Premises` and therefore inherits all its properties. For two other pairs of words – *walls* / *reception* и *walls* / *delegations,* one cannot find in the ontology such a natural connection.

7 NL Words and Semantic Elements

As mentioned above, the ontology serves as a semantic metalanguage, which means that SemSs are composed of the ontology elements – concepts, instances and properties (henceforth, semantic elements). All meaning bearing NL units are defined in terms of these elements. The correlation between NL words and semantic elements is far from being straightforward. Here several situations are possible. Let us look at some of them.

1. The simplest case: a NL word corresponds directly to an ontology element. For example, such words as *to win, to defeat, to lose, victory*, etc. all correspond to concept `WinEvent`.

2. A NL word does not have any ontological equivalent and simply disappears in SemS. In particular, this is the case of Lexical Functions of Oper-Func-Labor family. For example, sentence

(6) *Messi scored a goal*

has SemS

(6a) `hasAgent(GoalEvent, Messi)`

3. A NL word generates a fragment of SemS but it has no fixed ontological equivalent. Its interpretation depends on the context. Such words are particularly difficult for the semantic analysis. Let's consider the word *local*. In one of its meanings it denotes a place which is activated in the given context. For example, in the sentence

(7) *In Malaga we defeated the local team 2:1*

the *local team* is a team from Malaga. In the sentence

(8) *A deadly accident involved a local team at the Arizona Nationals*

the same phrase refers to a team from Arizona. In each case, the semantic analysis should discover which place is referred to by *local*. In sentences (7)-(8) above this place is directly mentioned in the sentence. However, this is not always the case. Sometimes the situation is more involved. The place referred to by *local* may be absent from the sentence altogether. Sentence

(9) *Charlemagne's remains were interred in the local cathedral*

is only appropriate if the place (the German city of Aachen) has already been activated in the previous text. Often, the place referred to by *local* is the place where the situation described in the sentence takes place.

(10) *The forum will allow local dealers to meet with the representatives of the manufacturer.*

Although this sentence doesn't provide any information on where the forum will take place, we know that the dealers are from the same place, and this fact can be used for inference. SemS of this sentence will contain the following fragment:

```
hasLocation(Forum01, Place01)
basedIn(Dealer03, Place01).
```

Let's give two examples more.

(11) *In the quarters Ahli Tripoli defeated Qatari 3:1. Their namesakes from Benghazi lost 0:3 to Saudi Arabia's Hilal.*

The first sentence contains a reference to the team whose name is *Ahli* and which is based in the city of Tripoli. The second sentence speaks of another team whose name is given by a reference to the first team by means of a shifter (*namesake*):

```
hasName(Team01, Ahli)
basedIn(Team01, Tripoli)
hasName(Team02, Ahli)
basedIn(Team01, Benghazi).
```

(12) *On May 28 Chelsea received Juventus. The visitors won 1:0.*

To understand where the match took place and who won, the system has to manipulate both semantic and encyclopaedic knowledge. First, it has to know that if a team receives another team, the match is played in a place which is the home for the first team but not for the second. This fact is accounted for by the semantic definition of *receive,* which is a rule that operates between NormSyntS y SemS:

Receives(X, Y) ⇔
```
hasParticipant(Match01, X)
hasParticipant(Match01, Y)
```

```
hasLocation(Match01, Place01)
basedIn(X, Place01)
basedIn(Y, Place02)
defferentFrom(Place01, Place02)
```

Besides that, there is a rule stating that a team is called *visitors* if it is not playing at home:

Visitors(X) ⇔
```
hasParticipant(Match01, X)
hasLocation(Match01, Place01)
basedIn(X, Place02)
defferentFrom(Place01, Place02)
```

Second, the system should possess the encyclopaedic information that Chelsea is based in London, and Juventus in Turin. With this information available, the system will construct SemS of the text and restore the following implicit information:

```
hasLocation(Match01, London)
hasWinner(Match01, Juventus)
```

4. An ontological concept does not correspond to any word of the sentence. This happens when the meaning is not expressed lexically but syntactically. For example, sentence (13) gives information on the match score but does not contain the word *score:*

(13) *FC Barcelona defeated Manchester United 3:1*

SemS of this sentence looks as follows:

```
hasWinner(WinEvent01, FC_Barcelona)
hasLoser(WinEvent01, Manchester_United)
inMatch(WinEvent01, Match01)
inMatch(MatchScore01, Match01)
hasValue1(MatchScore01, 3)
hasValue2(MatchScore01, 1)
inFavorOf(MatchScore01, FC_Barcelona)
```

8 Future research

As mentioned above, the project is at the initial stage. The coverage of the semantic rules is still small, although the syntactic parser, which constitutes an integral part of SemAnOn, has a very large coverage. In the near future we are planning to extend the coverage of the semantic analyzer to different types of constructions, in particular to the ones expressing intervals, temporal expressions and quantified noun phrases, to complete the football domain ontology and combine it with an upper/middle ontology, such as SUMO. Later, we hope to reformulate semantic restrictions in the subcategorization frames in terms of the ontology and use it widely for disambiguation tasks. Another direction of future research is connected with the construction of the fact repository.

Acknowledgements

The work reported in this paper has been partly supported by grants RFBR № 11-06-00405 and RSFH № 10-04-00040a which is gratefully acknowledged.

Bibliography

Apresjan, Ju. D. (2000). *Systematic Lexicography*. Oxford University Press, 2000, XVIII p., 304 p.

Apresjan, Jury, I. Boguslavsky, L. Iomdin, A. Lazursky, V.Sannikov, V.Sizov, L.Tsinman. (2003). ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. In: *MTT 2003, First International Conference on Meaning – Text Theory (June 16-18 2003).* Paris: Ecole Normale Supérieure, P. 279-288.

Apresjan, Yu., Tsinmam L. (2002). Formal'naja model' perifrazirovanija predlozhenij dlja sistem pererabotki tekstov na estestvennyx jazykax. *Russkij jazyk v nauchnom osveshenii,* № 2 (4), pp. 102-146.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American,* pp. 96-101, May 2001.

Boguslavsky, I., Grigoriev, N., Grigorieva, S., Kreidlin, L., Frid, N. (2000a). Dependency Treebank for Russian: Concept, Tools, Types of Information. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, , p. 987-991

Boguslavsky, I., Frid, N., Iomdin, L., Kreidlin, L., Sagalova, I., Sizov, V. (2002b). Creating a Universal Networking Module within an Advanced NLP system. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, vol. 1, 83-89.

Boguslavsky, I., Iomdin, L., Sizov, V., Timoshenko, S. (2010). Interfacing the Lexicon and the Ontology in a Semantic Analyzer. In: Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), Beijing, August 2010, pp. 67–76.

Clarke, J., D. Goldwasser, M. Chang and D. Roth. (2010). Driving Semantic Parsing from the World's Response. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010).*

Horrocks, I., P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof and M. Dean. (2004). SWRL: A Semantic Web Rule Language Com-bining OWL and RuleML. In: *W3C Member Submission 21 May 2004*.

Moldovan, D., Tatu, M., Clark, Ch. (2010). Role of Semantics in Question Answering. In: *Phillip C.-Y. Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, Lotfi A. Zadeh (Eds.) Semantic Computing, pp. 373-420.*

Nirenburg, S., and Raskin, V. (2004). *Ontological Semantics*. The MIT Press. Cambridge, Massachusetts. London, England.

Pall, Barney. (2006). POWERSET - Natural Language and the Semantic Web, invited lecture at *The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, (http://videolectures.net/barney_pell/)

Poon, H., & Domingos, P. (2009). Unsupervised semantic parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 EMNLP 09* (p. 1).

Ruifang Ge, Raymond J. Mooney. (2005). A Statistical Semantic Parser that Integrates Syntax and Semantics. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning.* Ann Arbor, MI, pp. 9--16, June 2005.

Towards the Annotation of Communicative Structure in Corpora

Alicia Burga(1), Simon Mille (1) and Leo Wanner (1,2)

(1) Department of Information and Communication Technologies
Pompeu Fabra University, C/ Roc Boronat, 138, 08018 Barcelona
(2) Catalan Institution for Research and Advanced Studies (ICREA)
<firstname>.<familyname>@upf.edu

Abstract

Communicative structure is central to the linguistic representation at nearly all levels of the Meaning-Text Models (MTMs). Its correlation with lexical and syntactic features makes it also essential for such natural language processing applications as text generation, which is about to undergo a significant shift from the symbolic, rule-based paradigm to the statistical paradigm. In the statistical paradigm, the availability of sufficiently large corpora annotated with linguistic information, and thus also with the communicative structure (CommStr), is critical. However, to the best of our knowledge, so far no corpora annotated with CommStr in the sense of the Meaning-Text Theory are available. We describe two experiments that explore how such corpora can be obtained. In the first experiment, a fragment of a Spanish Treebank is annotated manually. In the second experiment, we exploit the correlation of CommStr with syntactic features to annotate the English PropBank.

Keywords

communicative structure, MTT, treebank, annotation, corpus, Spanish, English

1 Introduction

Communicative Structure (CommStr) is central to the linguistic representation at nearly all levels of the Meaning-Text Models (MTM). As argued by Mel'čuk (2001), its various dimensions are signaled by lexical, syntactic, topological and prosodic means. For natural language processing (NLP) applications such as text generation (TG), especially the first two means are of relevance since they suggest that CommStr must be the driving instrument during lexicalization, i.e., mapping of semantemes to lexemes, and syntacticization, i.e., mapping of the Sem(antic) Str(ucture) onto the D(eep)-Synt(actic) Str(ucture) and of the DSyntStr onto the S(urface)SyntStr. Given that it is consensus among researchers that TG starts from abstract semantic or conceptual structures, one would expect CommStr to be of broad use in the field. However, this is not the case. The use of CommStr in TG is still rather

seldom. Only a restricted number of rule-based generators use it; see, e.g., (Iordanskaja et al., 1988; Wanner et al., 2003). Some works study the role of the CommStr for speech synthesis (Iomdin & Lobanov 2009; White et al., 2010; Iomdin et al. 2011), but, again, there are not many of them. This is certainly the reason why the CommStr has so far also been largely neglected in the recent statistical boom in NLP. As a consequence, hardly any corpus has been annotated with CommStr. We know just of the Prague Dependency Treebank, in which *Topic-Focus Articulation* (the equivalent of CommStr in the Prague School of linguistics) has been included (Hajič et al., 2006; Mikulová et al., 2006). This state of affairs is very unsatisfactory since our experience in statistical TG from semantic structures is that a corpus annotated with CommStr is indispensable (Bohnet et al., 2011). In what follows, we describe our ongoing work on the annotation of dependency treebanks of English and Spanish with CommStr. We explore how a large-scale annotation of the CommStr can be obtained: manually or drawing on treebanks that originally lack any communicative annotation. We believe that at least a part of the CommStr can be annotated automatically, based on the corresponding syntactic structure, but that the automatically obtained CommStr corpus should be completed by manual annotation in order to obtain fine-grained CommStrs suitable for machine-learning algorithms used in statistical NLP. The next section introduces, for convenience of the reader, the basics of the CommStr. Section 3 discusses our manual annotation exercises of the Spanish Treebank. Section 4 describes an experiment on the automatic derivation of some dimensions of CommStr from the syntactic and semantic annotation of the widely used English Treebank PropBank (Palmer et al., 2005), and Section 5 outlines our plans for future work in this area.

2 Basics of the Communicative Structure

In our interpretation of CommStr, we follow Mel'čuk (2001), who distinguishes eight communicative dimensions. The advantage of Mel'čuk's proposal is that (i) it is considerably more fine-grained than the other models of what is usually called *information structure*, and (ii) all of its dimensions are put in correlation with lexical, syntactic and prosodic means, while for information structure, only a correlation with word order and intonation has been discussed (Mikulová et al., 2006; White et al., 2010).

In what follows, we introduce the communicative dimensions that are of immediate relevance to TG and discuss how they are signaled by lexical and syntactic means particularly in the case of English and Spanish. We focus on the following five dimensions: 1. thematicity, 2. givenness, 3. focalization, 4. perspective, and 5. emphasis.[1] The dimension of **Thematicity** is given by the opposition between *rheme*, *theme* and *specifier*. *Rheme* is the content (or message) of the statement in question; *theme* marks what this message is about, and *specifier* sets the context of the message. In an English sentence, theme is, as a rule, expressed as the grammatical subject, while rheme is formed by the verbal governor with its object dependents and its local circumstantials. The sentential adverbials such as vocatives or sentential parentheticals form the specifier. **Givenness** captures the opposition between *given* and *new*. *Given* is the part of the statement that is known to the addressee, and *new* – the one that is

[1] We leave aside the dimensions of presuposedness, unitariness and locutionality because their role for generation still needs more reflection.

unknown. Gundel (1989) introduces four degrees of givenness, which correlate with different degrees of definiteness and pronominalization: *the*, *that*, *this*, and *it*. The new marker correlates with indefiniteness – *a*. **Focalization** marks parts of a statement that are in the focus of attention of the Speaker. The main means of focalization in syntax are dislocation, fronting, clefting, and conversion. **Perspective** (foregrounded vs. backgrounded vs. neutral) marks parts of the statement that are psychologically of primary / secondary relevance to the Speaker or that are not marked in terms of relevance. The main syntactic means to express *foregrounded* parts of a statement is raising; to express *backgrounded* parts, parenthetical constructions can be used. **Emphasis** deals with the emotive stress of parts of a statement. The main means to express emphasis are intonation and gestures; syntactic and lexical means include repetition (common, e.g., in Italian and Spanish) and special markers (such as the verbal *do* marker in English: *I do know what I am talking about*).

3 A first exercise in annotation of CommStr

The nature of the annotation of corpora with CommStr is different from the annotation with morphological and syntactic dependency tags since (i) CommStr tags need to be assigned to subgraphs or subtrees respectively rather than to single nodes or arcs, and (ii) the communicative tags are superimposed on the basic structure at a given level, i.e., SemStr, DSyntStr, SSyntStr, etc. In what follows, we focus on SemStr.

The first requirement for the annotation of the CommStr is to have access to the syntactic structure of the sentences to annotate. As mentioned above, in English and Spanish (as in most of the Indo-European languages), the syntactic structure directly reflects particular communicative features. In the following, we detail how to annotate the five communicative dimensions presented in Section 2. The examples cited have been gathered during the manual annotation of the CommStr on our multi-level annotated Spanish corpus. At this point, we annotated >400 sentences out of the 3.500 of the total corpus.[2] Consider Figure 1 for illustration.

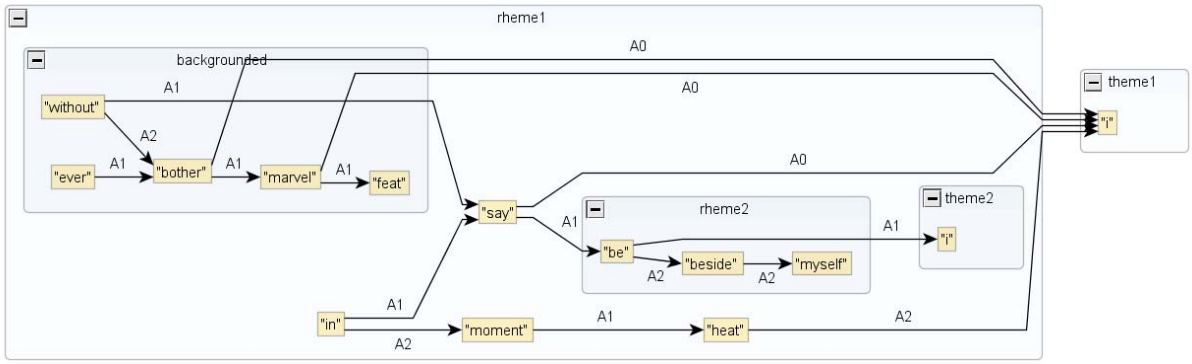


Figure 1: Sample CommStr annotation of *I've said in moments of heat, without ever bothering to marvel the feat, that I am beside myself.*

[2] For a preliminary presentation of the multilevel corpus, see (Mille et al., 2009), who use the same corpus as AnCora (Martí et al., 2008).

Thematicity: In the case of simple clauses, the annotation of thematicity is rather straightforward – except for the distinction between local and sentential adverbials, which can be problematic, especially if we talk about automatic annotation. However, thematicity is also recursive, which means that a subordinate clause (relative or completive) within a theme or rheme, and a specifier can contain their own theme and rheme. In order to differentiate the main communicative core from the embedded one(s), we use indices: theme$_{1/2/3}$…, rheme$_{1/2/3/…}$. Given that more than one embedded communicative structure is possible, the indices actually reflect the depth of each clause in the sentence.

The theme/rheme dimension can combine with any other communicative dimension, although no theme and rheme can be backgrounded as a whole. The governor of each theme and each rheme span is marked as the "main" communicative node (this information is particularly important when it comes to syntactically organize each communicative span during sentence generation).

Sentences containing indirect discourse deserve special attention due to their communicatively ambiguous interpretation: the subject is not necessarily theme nor is the main verb part of the rheme; rather, they can compose a specifier. This is the case when it is possible to replace the indirect discourse by *according to X*. Consider the sentence in (1) and its annotation. It could be interpreted – and consequently annotated – differently, if it were the case that the utterance is about the report and not about the stocks. Therefore, it is important to take into account the previous sentences (and even those that follow the sentence in question) when annotating manually, in order to evaluate whether or not the indirect discourse components are part of the thematicity core. Obviously, this kind of distinction would be extremely difficult to automate.

(1) [*El informe dijo que*]$_{Spec}$ [*las reservas de oro* […]]$_{T}$ [*eran de 18.300 millones de dólares*]$_{R}$
'The report said that the stocks of gold were 18.300 millions of dollars'.

According to Mel'čuk (2001), *wh*-words are necessarily rhematic. Following this assumption, *wh*-words are always annotated as part of the rheme. If the *wh*-word corresponds to the grammatical subject, the sentence is annotated as purely rhematic; cf. (2):

(2) ¿[*Quién* […] *puede haber diseñado una bacteria como la legionella* […]]?$_{R}$
'Who could have designed a bacterium as the legionella?'

The different components of a coordination are treated as being part of the same communicative component; see (3) – except for those cases where each component contains its own subject, as in (4).

(3) […] [*usted*]$_{T}$ [*es una persona poseída por el divino don de la caridad y quiere ayudar a sus semejantes* […]]$_{R}$
'You're a person obsessed with the divine gift of charity and want to help your fellow men.'
(4) [*El libro*]$_{T1}$ [*es divertido* […]]$_{R1}$, *y* [*su estilo*]$_{T2}$ [*un auténtico regalo*]$_{R2}$
'The book is fun and its style an authentic gift.'

In the corpus, we have found sentences with up to five themes and rhemes, although they correspond just to three levels of recursiveness; see (5) for illustration.

(5) [*Pero*]$_{Spec1}$ [*el juez*, [[*que*]$_{T2}$ […] [*dictaminó que* [*Microsoft*]$_{T5}$ [*incurrió en prácticas de monopolio*]$_{R5}$]$_{R2}$]$_{T1}$, [*opinó que* [*las medidas* […]]$_{T3}$ [*no son bastante severas*]$_{R3}$ *y se manifestó a favor de otro planteamiento* [*que*]$_{T4}$ [*dividiría a la empresa* […]]$_{R4}$]$_{R1}$
'But the judge, who ruled that Microsoft fell into monopoly practicing, expressed the view that the measures are not severe enough and declared himself in favour of another approach that would divide the firm'.

Givenness: This dimension can be marked very easily and directly: the nominal phrases that are introduced by definite articles (such as Sp. *el* and Engl. *the*) correspond to the first degree of givenness, demonstratives (such as Sp. *ese* and Engl. *that*) to the second degree, deictics (such as Sp. *este* and Engl. *this*) or possessive adjectives to the third degree. Nominal phrases headed by a pronoun are marked with the fourth degree of givenness. The rest of nominal phrases are signalled as new.

Obviously, there are no restrictions with respect to givenness when it comes to combine elements which express this dimension. In (6), the four different degrees of givenness are illustrated:

(6) a. […] *considera* [*estos cambios*]$_{G3}$ *lógicos, pues "resulta razonable que* [*las Bolsas europeas*]$_{G1}$ *unifiquen* [*sus normas de contratación*]$_{G3}$", *ya que* [*éstas*]$_{G4}$ *supondrán* ["*una ventaja* […]"]$_{N}$
'[He] considers logical these changes, since it seems reasonable that the European stock markets unify their transaction laws, given that they will mean "an advantage".'
b. […] [*la causa de* [*ese silbido*]$_{G2}$ o [*ese zumbido*]$_{G2}$]$_{G1}$ *es* [*la irritación d*[*el nervio acústico*]$_{G1}$]$_{G1}$
'The cause of that whistle or that buzzing is the irritation of the acoustic nerve.'
c. *¿No es acaso toda religión* [*la hipóstasis d*[*el conflicto entre* [*la inercia de* [*este mundo material*]$_{G3}$]$_{G1}$ *y* [*las supremas incitaciones de otro mundo*]$_{G1}$]$_{G1}$]$_{G1}$?
'Is it not every religion the hypostasis of the conflict between the inertia of this material world and the supreme incitements of another world?'

Given elements (of different degrees) can be included into other given elements (of the same or different degrees). It is also possible to find given elements within new elements (7), and vice versa (8).

(7) […] *Microsoft considera que* [*la separación de* [*la firma*]$_{G1}$]$_{G1}$ *es* [*un castigo demasiado duro para* [*las infracciones de las cuales se le acusa*]$_{G1}$]$_{N}$
'Microsoft considers the firm's division as a punishment too strong for the infractions for which it is accused.'
(8) […] *convencer a* [*la opinión pública*]$_{G1}$ *de* [*los riesgos de* [*un consumo*]$_{N}$ *y de* [*un crecimiento desbocado*]$_{N}$]$_{G1}$ […]
[…] 'to convince the public opinion of the risks of a consumption and a growth without control'.

Focalization: As mentioned in Section 2, the syntactic constructions that realize focalized content are quite obvious: fronted, cleft and promoted elements are marked as focalized.[3] In particular, adverbs and circumstantials that appear before the subject and object elements that appear before the governing verb are considered to be focalized. The rest of the sentence is marked as non-focalized (or neutral). During our annotation exercise, we have not found as

[3] The prominence of a focalized element with respect to the other elements is also reflected by intonation in that the nuclear accent is put over the focalized word (Hualde 2002). However, we do not delve into this issue here.

yet prototypical cases of dislocation or clefting, but we have found frequently cases of focalization through more subtle movements.

Focalization is directly linked to contrast. Thus, contrastive elements are focalized and marked as such; see (9) for illustration:

(9) […] *ha propuesto dar* [*a los fabricantes de ordenadores*]$_{\text{Foc}}$ *mayor flexibilidad* […] *y* [*a los consumidores*]$_{\text{Foc}}$ *más opciones* […].
'[He] has proposed to give to the computer manufacturers more flexibility and to the consumers more options.'

In order to differentiate focalized elements from foregrounded elements (which can also be marked through movement; see below), we have marked as focus those elements that contain by themselves a contrastive load:

(10) *Quiero* [*con esto*]$_{\text{Foc}}$ *decir que Medardo Fraile ha escrito un relato extraño y divertido*
lit. '[I] want with this to say that Medardo Fraile has written a strange and funny story.'

It is important to note that focalization is not recursive and focalized elements cannot appear within a backgrounded element. This information can be used when implementing a CommStr verification checker.

Perspective: Right-dislocated and parenthetical elements are marked as backgrounded. When circumstantial elements that normally appear in the periphery are located close to the verb and are not surrounded by commas, they are considered as foregrounded; other elements are marked as neutral.

Parenthetical elements are many times surrounded by commas. However, depending on the specific communicative situation, elements within commas can be interpreted as backgrounded or foregrounded. This is why we found difficulties when annotating manually this communicative dimension, and we had to turn to the sentences in the context to take a decision. When the context is of no help, we annotate by default those elements as backgrounded:

(11) *Austria conquistó 16 medallas en Salt Lake City* [*– 2 de oro , 4 de plata y 10 de bronce –*]$_{\text{Backgr}}$.
'Austria won 16 medals in Salt Lake City – 2 gold, 4 silver and 10 bronze.'

When clitics appear as markers of possession raising, they are considered foregrounded elements, equally to personal pronouns that appear before the *wh*-word in questions:

(12) […] ¿[*Tú*]$_{\text{Foregr}}$ *qué le regalarías por Reyes al duque de Feria?* […]
'What would you give to the Feria duke for Epiphany?'

Perspective is also recursive. Thus, it is possible to have foregrounded elements within backgrounded elements, as in (13), and vice versa, as in (14).

(13) *El gobernante*, [*con ganada fama* [*desde que llegó hace 16 meses al poder*]$_{\text{Foregr}}$ *de explotar al máximo su oratoria* […]]$_{\text{Backgr}}$, *enmudeció* […]
'The leader, with earned reputation since he got 16 months ago the power of exploiting the most his oratory, fell silent.'

(14) *Los últimos dos meses* […] *han estado marcados por insultos personales,* [*principalmente entre Labastida y Fox,* [*quienes se han dicho desde "mariquita", "feo"* […], [*entre otros calificativos*]$_{\text{Backgr}}$]$_{\text{Backgr}}$]$_{\text{Foregr}}$.
'The last two months have been marked by personal insults, especially between Labastida and Fox, who have said to each other from "wimp", "ugly", among other words.'

Appositions are another aspect that deserves discussion with respect to perspective. Even if appositions do not form backgrounded or foregrounded elements (Mel'čuk, personal communication), they do seem to play the role of a psychological relevance marker. We currently explore the precise nature of this marker.

Emphasis: As mentioned in Section 2, some lexical markers express an emotional load and thus emphasis. Although it is impossible to offer a comprehensive list of those markers (given that it is the context which finally defines whether or not an element is emphatic), we can make rough but useful generalizations. Thus, we annotate as emphatic some occurrences of the adverbial *una vez más* 'once again', *también* 'too', *muy* 'very', as well as superlative adjectives and adverbs (15). Which of them are in fact emphatic is decided upon the analysis of each occurrence.

(15) […] *procuraría que el regalo, además de* [*carísimo*]$_{\text{Emph}}$*, tuviera directamente que ver con el mayor vicio del obsequiado.*
'[I]'d try to make sure that the gift, as well as being expensive, it's also directly related to the greatest vice of who receives it.'

Repetition (approximate or exact, total or partial) of some elements is a syntactic means to emphasize a part of the utterance. We mark as emphasized the first element (in that we assume that it is emphasized through repetition) and as marker of emphasis the repetition itself:[4]

(16) *El libro es* [*divertido*]$_{\text{Emph}}$, [*muy divertido*]$_{\text{Emph\_Marker}}$ […].
'The book is fun, very fun.'

Emphasis is neither recursive nor obligatory. Even though theoretically emphasis can be combined with any other communicative dimension, so far we have found in the corpus emphasis combined with thematicity and givenness.

In addition to the five communicative dimensions discussed above, we annotate parts of utterances which appear within quotation marks at the surface with the "signalled" tag of the locutionality dimension.[5] Otherwise, it would not be possible to use quotation marks in statistical generation.

---

[4] We are using here the term 'repetition', but another more appropriate term could be proposed, given that sometimes the "repetition" consists on making explicit some semantic characteristics of the term to be emphasized, as in (i):
(i) […] [*titula*]$_{\textit{Emph}}$ *[en portada]*$_{\textit{Emph\_marker}}$ *"Villalonga normaliza las relaciones* […]*"*
'(It) heads in the front page "Villalonga normalizes the relations […]'

[5] According to Mel'čuk (2001), locutionality distinguishes between "communicating" and "signalling" utterances. The first ones pretend to explicitly communicate something and, in that sense, they can be headed by the phrase "I want you to know that…". The second ones just signal something that happens inside the speaker

4 Deriving CommStr from a Syntactic Dependency Annotation

The fact that part of the CommStr is signalled by lexical and syntactic means led us launch an experiment on the derivation of the CommStr from a syntactic dependency corpus annotation. The experiment has been performed on the dependency variant of the Penn Treebank (PTB) / PropBank (PB) corpus (Palmer et al., 2005), one of the most widely used corpora for English since it has been released in the standard corpus annotation format CoNLL (Hajič 2009), which contains in the same data structure syntactic and semantic annotations.

4.1 Experiment of the derivation of CommStr

The automatic derivation of the CommStr from the PTB/PB annotations has been performed using the rule-based MATE graph transducer (Bohnet & Wanner, 2010). The derivation is based on a set of rules which use semantic and superficial syntactic and topological criteria available in PB and PTB. For instance, the subject of a sentence is marked as theme and the corresponding VP as rheme; an indefinite NP is marked as new; and so on.

The available criteria are, obviously, too crude and too simplistic to capture the information structure in its entirety. The fact that the CommStr–SyntStr projection is not isomorphic makes the derivation even more difficult. Therefore, we focused on the derivation of the opposition theme vs. rheme (ignoring the feature of Specifier), the dimension of Givenness and a combined version of Perspective and Focalization we called "Foregroundedness".

4.2 Assessment of the derivation

Despite the limited range of criteria we could use for the derivation of the partial CommStr, the evaluation below shows that the obtained CommStr may well serve as a first approximative annotation.

4.2.1 Quantitative Evaluation

To assess the quality of the derivation, we performed a quantitative evaluation in which we compared the automatically obtained CommStr with a gold standard of 90 sentences.[6] Table 1 shows the results of our quantitative evaluation:'thematicity *p*/*r*' stands for precision and recall of the theme/rheme introduction; 'main *p*/*r*' for precision and recall of the marking of the main node of all themes and rhemes, and 'th-rh pairs *p*/*r*' for precision and recall of the identification of the theme/rheme alignment (each theme is marked as being the theme of a particular rheme of the sentence). 'foregr/backgr *p*/r' stands for the accuracy of the perspective annotation (foregrounded/focalized vs. backgrounded vs. neutral), 'depth *p*/*r*' for

(they do not express linguistically the communication act), and in that sense they cannot be negated or questioned.

[6] We are fully aware that 90 sentences are not sufficient to objectively assess the results of our experiment. However, even with such a small gold standard corpus it is possible to estimate whether the adopted strategy is promising or not.

precision and recall of the recursive theme/rheme (primary, secondary, terciary, etc.) annotation, and 'given *p*/*r*' for the accuracy of givenness annotation.

thematicity		main		th-rh pairs		foregr/backgr		depth		given	
p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
0.986	1.0	1.0	0.951	0.914	1.0	0.905	0.358	0.807	1.0	1.0	0.986

Table 1: Precision and recall for the automatic introduction of CommStr dimensions

The numbers show that the identification of the main node, theme/rheme and givenness works well. This is because these notions actually correlate very much with some prominent syntactic features that could be deduced from the PTB annotation. The accuracy of the annotation of the recursive theme/rheme structure is somewhat lower. This is because every node receives its thematicity feature from the main node of the span it belongs to, but each node can have more than one governor, and each governor can belong to a different communicative span. The recall of the assignment of the perspective is rather low (0.358), although the precision is high (0.905). This means that syntactic and topological clues only are by far not sufficient to determine which element is to be marked as foregrounded, which one as backgrounded, and which one is neutral with respect to communicative prominence. Other types of clues are also needed.

thematicity			main			th-rh pairs			foregr/backgr			depth			given		
tp	fp	fn	tp	fp	fn	tp	fp	fn	tp	fp	fn	tp	fp	fn	tp	fp	fn
704	10	0	135	0	7	64	6	0	38	4	68	569	136	0	70	0	1

Table 1: Total numbers of the quality of the annotation of the individual communicative features ('tp' stands for "true positives", 'fp' for "false positives", and 'fn" for "false negatives")

4.2.2 Limitations of the derivation of CommStr from syntactic annotation

The automatic derivation of CommStr from a syntactic annotation can only be partial. First of all, in Indo-European languages, there are few distinctive syntactic constructions for focalization, mainly clefting and left dislocation. However, a left dislocation can be difficult to interpret, since it can also correspond to neutrality from the point of view of focalization (consider, e.g., *yesterday* in *Yesterday, I went to the beach.*). Emphasis is often spotted thanks to the presence of particular cue words in a particular position. For perspective, the presence of parentheses is a clear marker of backgroundedness, but as far as the other features are concerned, it is necessary to look at the positioning of the groups. The fact that importance is given to the ordering among the components of a sentence is also a problem by itself: it raises issues when it comes to languages with free word order, such as Russian, for instance. In addition, some reasoning is necessary in order to interpret a sentence; an algorithm will probably never be able to recognize a slightly peculiar construction or a combination of words which are intended to signal a particular communicative goal of the speaker.

5 Conclusions and future work

The rule-based derivation of CommStr from syntactic annotation such as PennTreeBank and PropBank is an option that can be considered to ensure a short term availability of a corpus annotated with CommStr. However, if a high quality, detailed annotation is targeted, machine learning (ML) based annotation seems more adequate. Given that ML-based annotation requires manually annotated corpora as training material, we need to enlarge our corpora, as well as to guarantee the quality of the annotation compiling precise guidelines for the annotators, using such metrics as inter-annotator agreement and foreseeing a posterior revision iteration. But in order to be able to compile precise annotation guidelines we still need to discuss and decide how to treat certain phenomena, such as the distinction between focalized and foregrounded elements, or the definition of emphatic markers.

Acknowledgements

Thanks to Bernd Bohnet for his help with the derivation of CommStr from the PropBank annotation and the subsequent evaluation. This work has been partially supported by the European Commission under the contract number FP7-ICT-248594, by the Spanish Ministry of Science and Innovation under the contract number FFI2008-06479-C02-02 and by the Funds FEDER of the European Commission.

Bibliography

Bonhet, B. & L. Wanner. 2010. Open Source Graph Transducer and Interpreter Development Environment. In Proceedings of the LREC.

Bonhet, B., L. Wanner, & S. Mille. 2011. Statistical Language Generation from Semantic Structures. In Proceedings of the DepLing.

Gundel, J., N. Hedberg, & R. Zacharski. 1989. Giveness, Implicature and Demonstrative Expressions in English Discourse. In CLS-25, Part II (Parasession on Language in Context), pages 89–103. Chicago Linguistics Society.

Hajič, J. et al. 2006. *Prague Dependency Treebank 2.0* . In Linguistic Data Consortium, Philadelphia.

Hajič, J. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Proceedings of the CoNLL.

Hualde, J.I. 2002. Intonation in Spanish and the other Ibero-Romance languages: Overview and status quaestionis. In *Romance Phonology and Variation. Selected Papers from the 30th Linguistic Symposium on Romance Languages, Gainesville, Florida, February 2000*, ed. by Caroline Wiltshire and Joaquim Camps, 101-115. Amsterdam: Benjamins.

Iomdin, L.L. & B.M. Lobanov. 2009. Syntactic Correlates of Prosodically Marked Elements of the Sentence and Their Role in the Tasks of Text-to-Text Speech Synthesis. In Proceedings of the Dialog '09 Conference.

Iomdin, L.L., B.M. Lobanov & Ju.S. Gecevich. 2011. The Talking ETAP. Using the ETAP Parser in Russian Speech Synthesis. In Proceedings of the Dialog '11 Conference.

Iordanskaja, L. N., R. Kittredge & A. Polguère. 1988. Implementing a Meaning-Text Model for Language Generation. In *Proceedings of COLING 1988*.

Martí, M.A., M. Taulé, L. Márquez, & M. Bertran. 2008. Ancora: A Multilingual and Multilevel Annotated Corpus, Pending to be published (http://clic.ub.edu/corpus/ancora-publicacions)

Mikulová, M. et al. 2006. *Annotation on the tectogrammatical level in the Prague Dependency Treebank. Reference Book.* http://ufal.mff.cuni.cz/pdt2.0update/doc/tr-ref-cz-en.pdf

Mel'čuk, I.A. 2001. *Communicative Organization in Natural Language : The Semantic-Communicative Structure of Sentences.* John Benjamins Publishing, Philadelphia.

Mille, S., Burga, A., Vidal, V. & Wanner, L. 2009. "Towards a Rich Dependency Annotation of Spanish Corpora". In *Proceedings of SEPLN'*09, San Sebastian.

Palmer, Martha, D. Gildea, & P. Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles, in *Computational Linguistics Journal*, 31:1.

Wanner, L, B. Bohnet, & M. Giereth. 2003 Deriving the Communicative Structure in Applied NLG. in *Proceedings of the 9th European Natural Language Generation Workshop* at the Annual Meeting of the Association for Computational Linguistics, Budapest, 111-118.

White, M, R.A.J. Clark, & J. Moore. 2010 Generating tailored, comparative descriptions with contextually appropriate intonation. *Computational Linguistics*, 36(2):159–201.

English/Russian to UNL Enconverter[1]

Viacheslav Dikonov

Computational Linguistics Lab – IITP RAS
Bolshoy Karetny per. 19, Moscow, 127994, Russia
dikonov@iitp.ru

Abstract

This paper presents a new UNL enconverter - an automatic rule based tool to transform English and Russian text into semantic graphs. It is built on top of the ETAP-3 linguistic processor, which is largely based on the general linguistic framework of the Meaning↔Text theory by Mel'čuk. We describe the principles and operation of the enconverter and highlight its key features. The UNL graphs produced by this system are understood by several UNL deconverters which transform them back into text in different languages. The system represents a step further beyond the level of deep syntactic structure towards building a complete Text→Meaning model.

Keywords

UNL, ETAP, semantic graph, semantic representation, rule based system, interlingua

1 Introduction - What is UNL and UNL enconverter?

UNL (*Universal Networking Language*) is an interlingua based on the semantic graph formalism. It is designed to capture the meaning of text in any natural language with good precision, store it in the computer memory and support generation of equivalent text in any other natural language. The semantic UNL graphs are both machine and human readable and can be post-edited to correct any errors introduced by automatic analysis. A *UNL enconverter* is a software instrument to analyze a natural language text and represent its meaning as a UNL graph. The reverse transformation is performed by *UNL deconverters*.

A UNL graph consists of nodes linked with directional semantic links of several broadly defined types, such as "agent", "partner", "posessor", "place", "purpose" etc. A node may have multiple incoming edges. The nodes may contain either single concepts or other UNL graphs. Nodes of the latter kind are called *hypernodes*.

[1] This work received finacial support from RFBR (Grant 08-06-00367).

The basic lexical units of the UNL language are *concepts* usually representing a sense of a word or an idiom of some natural language (not necessarily English) as distinguished by explanatory dictionaries. There are concepts corresponding to abstract ontological terms as well, e.g. "*thing*", "*do*", "*abstract_thing*". Each concept has at least one[2] label called *UW* (*Universal Word*). Each UW expresses exactly one concept. The idea is that all concepts of all languages should receive unique UWs. If UNL lacks a UW for some concept, it can be added on demand. A UW consists of a headword and a list or constraints clarifying its meaning, e.g. *drink(icl>consume>do,equ>imbibe,agt>person,obj>matter)*. The headwords are usually English words, but words of other languages in Latin transcription may be used too. The constraints are UNL relations linking the concept to other concepts.

The graph nodes can carry special attached marks called *attributes*. UNL attributes are used to encode common grammar categories, subjective standing of the speaker towards the spoken, logical relations, etc. There is a finite set of possible attributes. The names of all UNL attributes start with the sign @, e.g. @past, @unreal, @topic, @polite, @possibility...

UNL graphs resemble both deep syntactic and semantic structures described by Mel'čuk (Mel'čuk, 1974). Table 1 summarizes key differences between UNL and SemStr.

SemStr	**UNL graph**
Semantemes can be decomposed into graphs consisting of elementary semes.	UNL does not decompose its concepts into any smaller units.
Graph edges in SemStr are nameless or numbered. They can be interpreted as semantic links only after consulting the dictionary entries of the predicates.	All graph edges must be labeled with one of 46 predefined semantic types. The resulting links retain their meaning even when the lexical contents of the nodes is masked.
Links are always directed from predicates to objects.	UNL still has the syntactic *mod* (modifier) relation which goes from an object to the modifying predicate. The inverse relation *aoj* (thing with an attribute) is also available.
A regular SemStr should contain only nodes and edges.	UNL graph nodes can have attributes. Some attributes, e.g. modal ones, can be expanded into additional nodes and edges. Another characteristic formal feature is the use of hypernodes.
SemStr should include discourse indicators as « metapredicates ».	Attributes can serve as indicators of topic and emphasis, but there is no established way to indicate theme and rheme. In order to preserve the sequential order of sentences within a text, they are represented as separate numbered graphs.

Table 1: differences between SemStr and UNL graphs

2 What is inside the UNL enconverter?

Our UNL enconverter operates on top of the ETAP-3 syntactic parser and uses the ETAP-3 (Apresjan et. al 1992) software platform, its dictionary and rule formats. It consists of the ETAP UNL dictionary and a set of rules to produce UNL graphs.

[2] It is unrealistic to say that any concept will always have just one associated UW, because there are multiple UNL-related projects. UWs may be created and edited by different people and even after merging all equivalent UWs together in a common dictionary alternative forms still have to be recognized in legacy documents encoded using older versions of UNL.

2.1 UNL Dictionary

The current UNL dictionary in ETAP, which is directly used by our enconverter contains over 82 000 UNL concepts linked with over 47 000 English and 22 000 Russian words. It is a sub-set of the general UNL dictionary of concepts, expressed in the native ETAP format. It includes only the UWs that have direct association with the words registered in the English combinatorial dictionary of ETAP.

The general UNL dictionary of concepts is larger and is currently under active development. It already includes some data currently unused by ETAP, such as French-UNL dictionary and a large ontology. The structure and principles of the general UNL dictionary are described in (Boguslavsky & Dikonov, 2008). The resource is proposed as a common standard for several research groups, which joined into the so-called "U++ Consortium", namely the Russian group at IITP (Moscow), the French group at CLIPS-IMAG (Grenoble), the Spanish group at UPM (Madrid) and the Indian group at IIT (Mumbai). The UWs in the common dictionary are also linked with Princeton Wordnet synsets and terms of the SUMO ontology. It is an open resource which includes data from other open resources and our plan is make it available for the public under GPL and CC licenses.

2.2 UNL en-conversion pipeline

In order to convert a text into UNL graphs ETAP splits it into sentences and performs full syntactic parsing of each sentence first. The result is a dependency tree closely following the Mel'čhuk's formalisms. It should be mentioned, that our approach to UNL en-conversion is totally based on dependency trees . The UNL en-conversion process begins after obtaining the syntactic parse. It includes three well defined stages:

- Preliminary markup and light normalization according to UNL rules,
- Lexical disambiguation,
- Semantization of syntactic relations & creation of hypernodes.

Each of these steps is served by dedicated sets of rules and will be described further.

Perhaps, the best way to explain the operation of any natural language processing system is to show how it would process a real sentence step by step. Let us take a random example from a news article about the E.Coli epidemic in Germany (1) "*They said that despite nearly 200 new cases in Germany - the centre of the outbreak - infection rates were dropping.*" and describe all intermediate states and its internal representation in the system.

2.2.1 Syntactic tree

First, ETAP produces a single dependency tree. The tree undergoes normalization, which turns it into the so-called *Normalized syntactic structure* (*NormSS*) used by ETAP for transfer-type translation into other languages.

The UNL enconverter takes English NormSS as its input. If we choose to analyze Russian, the Russian NormSS will be converted into the English NormSS before enconversion. The

inevitable losses caused by such conversion are partially offset by storing the default UNL equivalent (the most common sense) of Russian words without direct English equivalents in the special internal field for default translations. Figure 1 shows the initial stage of enconversion.

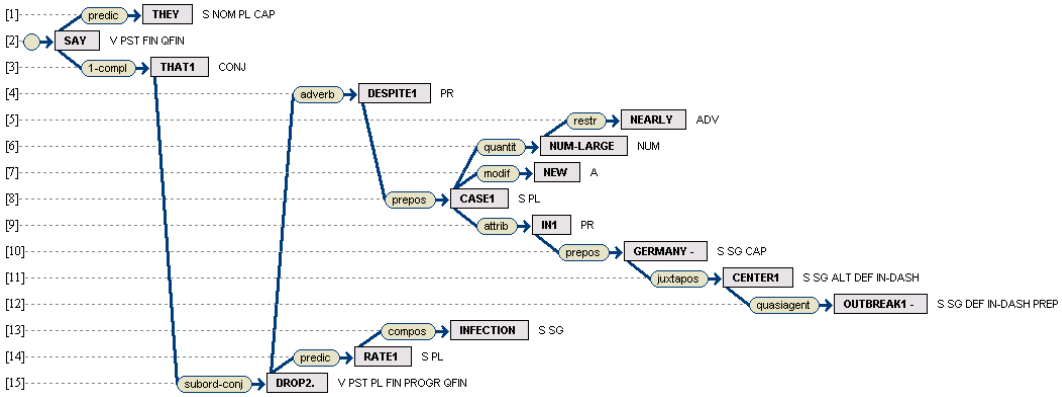


Figure 1: Normalized syntactic structure of example 1 before conversion.

2.3 Preliminary markup

The first step of preliminary markup serves to reduce complexity of further processing and avoid certain errors by inserting UNL attributes that will guide other rules at further stages. The overall structure of the syntactic tree and contents of its nodes usually remain unchanged.

There are several important tasks performed at this stage:

- The system detects if the current sentence contains questions, addresses or imperatives and adds appropriate UNL attributes.
- Tense, aspect and actuality information is analysed and UNL attributes of tense (@past, @present, @future), aspect (@complete, @progress, etc.) and unreality (@unreal) are inserted. If the sentence contain verbs should/would/could/might, special rules are used to determine whether they express tense, subjunctive mood or modality to prevent false interpretations.
- If the sentence contains passive constructions, the UNL attribute @topic is inserted to mark the real object of the action and the passive construction is converted to active by replacing "predic" links with "1-compl".
- Future head nodes of the graph and some of its hypernodes are marked. This rule finds the main predicate and all of its coordinated predicates and inserts special internally used attributes that request (@_HN) or prohibit (@_HNTOP) the creation of hypernodes with the marked predicates as their heads. It prevents putting the whole sentence into a hypernode.
- There are also rules to markup the semantics of participles. The phrases like *"killed people"* and *"drunk people"* are syntactically similar, but the killed are patients of killing while the drunk are agents of drinking. This fact must be reflected in the UNL graph as *killed* $-^{obj}\rightarrow$ *people* and *drunk* $\leftarrow^{mod}-$ *people*. Likewise, present participles can denote either states or action. For example *"the boy living in Sweden"* produces *live* $\leftarrow^{mod}-$ *boy* and *"the boy driving the car"* becomes *drive* $-^{agt}\rightarrow$ *boy*.

- Certain anaphoric constructions are detected and a proposed special relation "*ref*" is created between the anaphoric word and its antecedent. Here are some examples: *"Tom with his toy"* (he→Tom) *"Jane and her bag"* (Jane←her) *"Tom and Tom's books"* (Tom[1] = Tom[2]) *"Sir, i brought your dinner"/"Your dinner, sir"* (Sir←you) *"This is for you, my dear."* (you→dear). It is impossible to detect any anaphoras which span several sentences because ETAP processes just one sentence at a time and does not store any information about previous sentences.

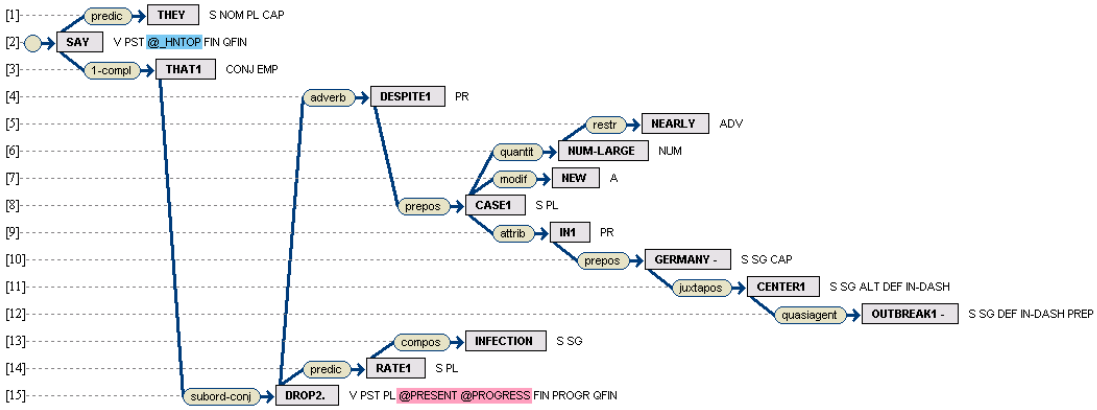


Figure 2: Preliminary markup of example 1.

In our example sentence (Fig.2 above) we see only two changes. The service attribute @_HNTOP (in blue) marks the main predicate of the sentence and prevents it from becoming the starting point of an unnecessary hypernode that would wrap the whole sentence. Other two UNL attributes @present and @progress show that the time of rate dropping was present relative to the moment of speech and this process has not ended by that moment.

2.4 Lexical disambiguation

The goal of this step is to replace all words in the tree with suitable UWs. No syntactic links are changed here and the structure still remains a tree (See Fig.4). Each word can be associated with several UWs representing its different senses. The choice of the correct word sense and UW is important for further processing beause UWs provide access to semantic information in the UNL dictionary. It includes ontological classification and the data necessary to interpret completive syntactic links in terms of UNL relations.

Lexical disambiguation is a classical problem, which has no good solution so far. The ETAP platform offers just three possible ways to do it without turning to external tools. First, all UWs are ranged according to their frequency measured against the Semcor corpus. It helps to improve the number of correct guesses by taking the first – most frequent – sense in the list. Second, there is a system of rules that analyze the immediate tree context and automatically insert the right UW if the context matches a known scheme. Finally, the system has an interactive mode, shown in Fig.3, which allows to ask the user about the true meaning of the words in the analyzed sentence. This method achieves high precision but it is time consuming and tedious for the user.

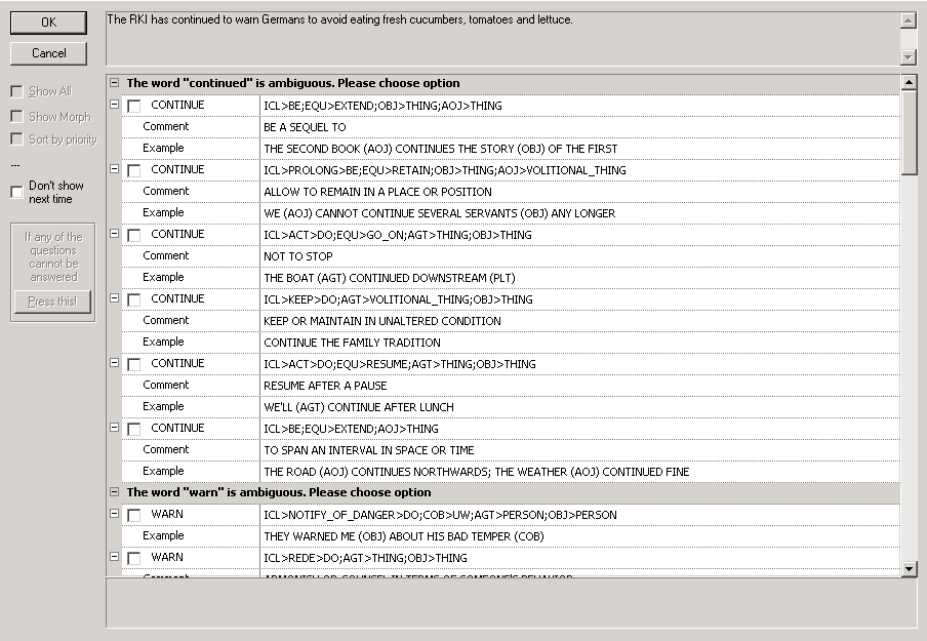


Figure 3: Interactive disambiguation for UNL.

The efficiency of the interactive disambiguation can be improved by changing the user interface so, that the user would see the sentence preannotated with automatically chosen UWs and change only the false choices by indicating one or more suitable ones. Unfortunately, this has not been implemented.

One of the major problems with disambiguation rules is that there must be a set of rules for each word sense describing its possible contexts. Their number could easily go over a million. To mitigate this problem we use a limited set of less than 40 very simple generalized macro rules matching small predefined fragments of the NormSS tree. The fragments may include the parent and 1 – 3 dependent words with their morphosyntactic and semantic features and relations. The actual words, names of relations and features are substituted by variables. The exact values of the variables depend on the currently analyzed word and are supplied by the dictionary. Let us have a look at one of the simplest rules, called UNL-CONV1.15:

REG:UNL-CONV1.15 TRANSLATION DEPENDING ON FEATURES OF A DEPENDENT LEXEME //ВД
N:1 *Subrule number 1*
CHECK ***Conditions that must be satisfied in order to apply the rule***
1.1 DOM-EQU(X,,T1,T2)* *Current node X commands another node * with feature T2 via relation T1.*
2.1 PARAMSYS(DISAMBMODE,DISAMB_LEX,DISAMB_LEXSYNT) *Do not override human choice in interactive mode.*
2.2 =(X,EMP) *Skip nodes marked as semantically empty.*
DO ***Actions performed if all conditions are satisfied (This part is the same for all such rules.)***
1 ZAMUZ:X(UNL,LUNL) *Replace the word in node X with the UW given by the LUNL variable.*

In plain language it reads as follows: "*If the current word has a syntactic daughter linked by relation T1 and this daughter has syntactic or semantic feature T2, choose the UW specified by LUNL*". The verb "DROP" has the following records in the dictionary:

...
TRAF:UNL-CONV1.15 (reference to the rule template shown above)
T1:3-COMPL,T2:'OBJECT',LUNL:DROP{ICL>MOVE>BE;EQU>DESCEND;PLT>THING;PLF>THING;AOJ>CONCRETE_THING}
T1:3-COMPL,T2:'HUMAN',LUNL:DROP{ICL>UTTER>DO;AGT>PERSON;OBJ>ABSTRACT_THING;REC>PERSON}
T1:PREDIC,T2:'PARAMETER',LUNL:DROP{ICL>CHANGE>OCCUR;OBJ>THING} *"RATES ARE DROPPING"*
...

The last line causes the system to interpret "drop" as "abrupt decreasing" in the context of a parameter. Such rule references come in sequence ranged from more complex rules with more

parameters to simpler ones with less parameters. This forms a decision tree that can be compared with sorting of diamonds, where there are many sieves with progressively bigger holes. Smaller gems (better defined contexts) fall through the first ones and only the largest reach the last sieve (the most general context). It is also possible to specify the second best answer for each defined context by setting a special flag "OPT:1". The user can manually ask the system to produce a different version of the graph automatically and the alternative outputs will contain such second answers. If there are no rules or none of them provides an answer, the first entry in the list of UWs associated with the current word will be chosen.

Although this mechanism may seem to be quite old-schoolish, it has been designed with the possibility in mind to use trendy statistical methods to generate massive amounts of such rule entries by analyzing a corpus.

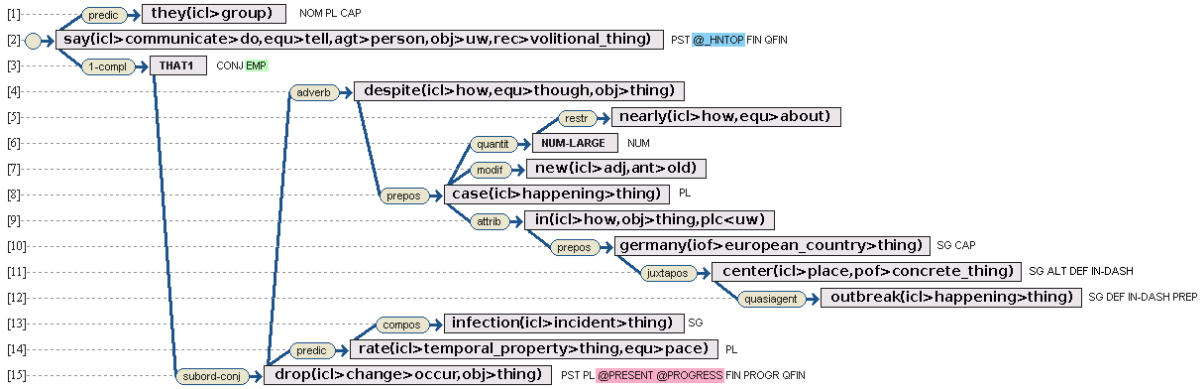


Figure 4: Lexical disambiguation of the example sentence 1.

In Figure 4 there are two words that have not received any UW: the number "200" and "THAT1", which is marked as semantically empty by the EMP attribute. All words with this mark will be deleted at the next stage. Prepositions, such as "IN" in our example, and conjunctions are assigned lexical meanings too.

2.5 Semantic interpretation of syntactic links

This step is dedicated to the actual transformation of the syntactic tree into a true semantic graph (See Fig.5 below).

[S:00]
{org:en}
They said that despite nearly 200 new cases in Germany - the centre of the outbreak - infection rates were dropping.
{/org}
{unl}
agt(say(icl>communicate>do,equ>tell,agt>person,obj>uw,rec>volitional_thing).@entry.@past,they(icl>group).@pl)
man:02(drop(icl>change>occur,obj>thing).@entry.@pl.@present.@progress,despite(icl>how,equ>though,obj>thing))
man:02(200,nearly(icl>how,equ>about))
qua:02(case(icl>happening>thing).@pl,200)
mod:02(case(icl>happening>thing).@pl,new(icl>adj,ant>old))
obj:02(despite(icl>how,equ>though,obj>thing),case(icl>happening>thing).@pl)
plc:02(case(icl>happening>thing).@pl,germany(iof>european_country>thing))
obj:01(center(icl>place,pof>concrete_thing).@entry.@def,outbreak(icl>happening>thing).@def)
mod:02(rate(icl>temporal_property>thing,equ>pace).@pl,infection(icl>incident>thing))
obj:02(drop(icl>change>occur,obj>thing).@entry.@pl.@present.@progress,rate(icl>temporal_property>thing,equ>pace).@pl)
cnt:02(germany(iof>european_country>thing),:01.@dash)
obj(say(icl>communicate>do,equ>tell,agt>person,obj>uw,rec>volitional_thing).@entry.@past,:02)
{/unl}
[/S]

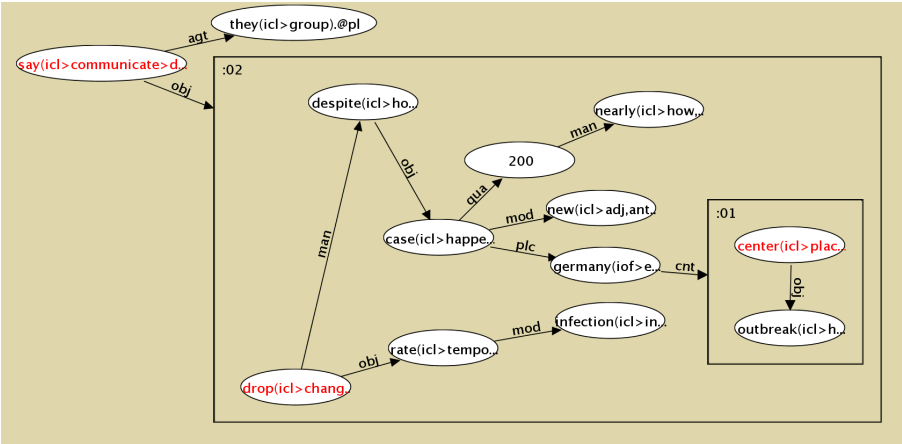


Figure 5: The resulting graph and UNL code of the example sentence 1.

At this stage the structure may cease to be a tree. All syntactic links except the emergency "fictit" link must be replaced by UNL relations. There are two tiers of rules that do it: the macro rules, which perform transformations specific to the chosen UWs, and general rules. First, the macro rules replace 1,2,3,4-completive and other links representing the government patterns of predicates with corresponding semantic relations specific to each UW. The links around prepositions and conjunctions are treated in the same way depending of the chosen meaning. Second, the general rules interpret remaining syntactic links and replace syntactic features with corresponding UNL attributes. Finally, the hypernodes are created where necessary. Figure 5 shows the final result of the conversion. This particular graph contains two hypernodes. Node :02 describes the contents of what has been said and :01 contains the inserted side note about Gemany "*the center of the outbreak*" and holds the attribute @dash to indicate original formatting. The empty word THAT has been deleted and the preposition IN in its spatial meaning was reduced to the equivalent relation "plc" (place). All information about the original word order is erased. The red nodes at the graph's root and in each hypernode are the starting points for graph interpretation.

3 Highlighted features

The UNL enconverter project introduces some important new features and resources both to ETAP and UNL. The most important new features in ETAP were:

- interlingua-based translation and potential support for extra language pairs via UNL deconverters built by other groups,
- support for hypernodes and non-tree graph structures,
- ability to modify the rule behavior depending on the system's operating mode,
- a new dictionary with extra semantic information.

Most important contributions to the UNL development are:

- a converter for English and Russian supporting latest specifications and proposals,
- common dictionary of UNL concepts, which is already used by the Russian and French groups and includes UNL↔French and UNL↔Russian dictionaries,
- an improved system of UNL attributes to encode modality and evidentiality.

3.1 Non-tree structures

There are several cases when the enconverter can create non-tree structures. 1) The UNL relations *and*, *or* and argument relations of participles are reversed, i.e. go from syntactic daughters to their parent nodes. In some sentences, as in Fig. 6, this results in non-tree graphs.

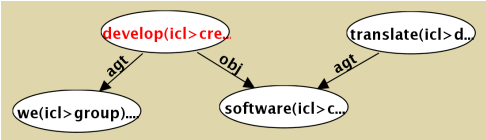


Figure 6: Non-tree graph of the phrase "We develop translating software."

2) The semantic scopes (Boguslavskij, 1996) of modifiers syntactically attached to one member of a coordinated chain but covering other members as well can be represented either by using a hypernode or several UNL links connecting the modifier with each of the modified chain members. For example, the scope of modifier "*fresh*" in (2) "*The RKI has continued to warn Germans to avoid eating fresh cucumbers, tomatoes and lettuce.*" includes all three members of the chain "*cucumbers, tomatoes and lettuce*" (See Fig.7).

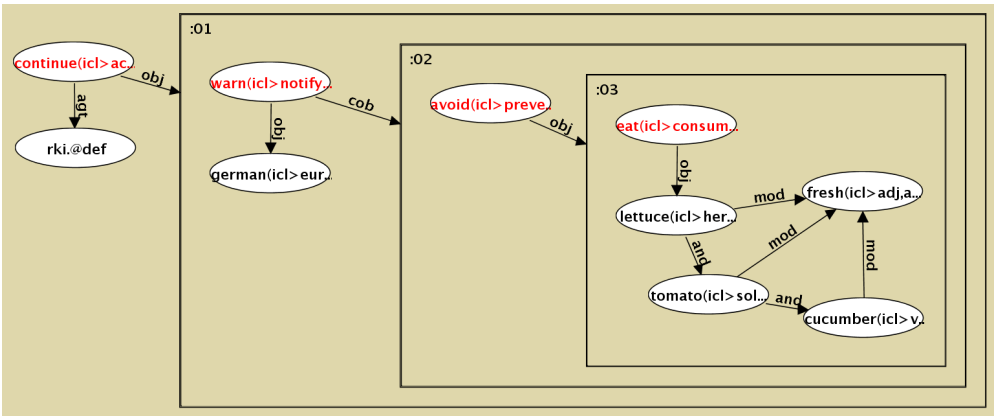


Figure 7: Non-tree graph of the phrase "The RKI has continued to warn Germans to avoid eating fresh cucumbers, (fresh) tomatoes and (fresh) lettuce."

3) Referential relations[3] by their nature disregard the tree structure. An anaphoric word and its antecedent or two different namings of the same object are always placed in different subtrees. Connecting them in means creation of a non-tree graph. For example, the sentence (3) "*The cat eats the mouse the cat catches*" discussed by Etienne Blanc (Blanc, 2005) contains two mentions of the same cat and should be represented in UNL by a circle graph, where both "*cat*" and "*mouse*" receive two parent predicates "*eat*" and "*catch*" (See Fig.8) while (4) "*Tom and his toy*" contains the proposed optional *ref* link between "*he*" and "*Tom*".

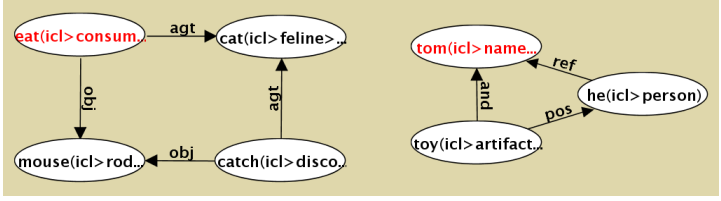


Figure 8: Non-tree graphs 3 and 4 are caused by coreference.

[3] We argue that using numeric indices to show coreference in UNL is not enough and we must introduce an extra UNL relation "ref" (referential) to support anaphoras.

3.2 Hypernodes

In UNL hypernodes are also called "scopes". They have three different uses: 1) to represent the semantic scopes of predicates, which are broader than one word, 2) to isolate fragments of text, such as quotations, remarks, etc., inside quotes, brackets and double dashes and carry the respective attributes of formatting, 3) to segment a complex sentence into smaller parts containing individual propositions with a single predicate. Such segmentation of sentences is characteristic of our enconverter. It helps the user to view and understand large complex UNL graphs in their visual form. More about it is written in (Dikonov, 2008). The hypernodes in figures 5, 6, 9 well demonstrate the rule of segmentation and treatment of double punctuation marks (node :01 in figure 5). Hypernodes constitute a major addition to the ETAP-3 formalism which extends previous limits on the kind of supported structures.

3.3 Modality in UNL

The new enconverter supports an improved system of UNL modal attributes. It is described in detail in a separate paper (Dikonov, 2009)[4]. Its development was stimulated by the drawbacks of the old set of attributes prescribed by the (UNL specifications, 2005). The old attributes remained ambiguous and were too closely tied to the English modal verbs, which made encoding of some modal utterances in other languages inadequate. Besides, they lacked mutual organization that would enable proper translation of modals between languages having not fully compatible systems of modality. For example, some languages enforce the distinction between being physically capable of doing something and knowing how to do it. They require different modal words where English uses the same modal verb "can".

Figure 9 shows the result of automatic parsing of the English sentence (5) "*Could you be so kind as to advise me of what could have happened?*" which demonstrates two meanings of "*could*" and the ability of the enconverter to disambiguate modal expressions and detect formulas of politeness.

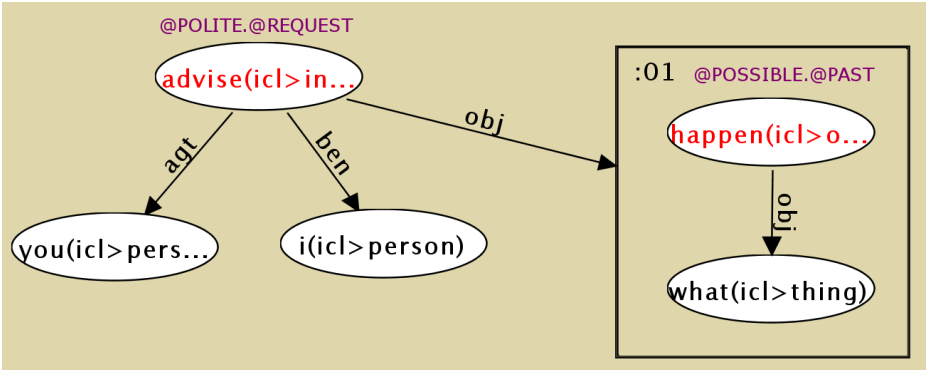


Figure 9: Disambiguation of modal verbs.

The new system of modal attributes is backwards compatible with the old one. Each UNL deconverter supporting the new attributes will be able process older graphs without any modifications because all old attributes are mapped to new ones.

[4] An extended version in English is currently available on request.

Current status and problems

Our UNL enconverter is able to produce semantic representation of English and Russian text but there are still things to improve. We are not able to give precision and recall scores against a large manually checked UNL corpus because such corpus has yet to be prepared. There are two problems that cause errors: wrong syntactic structures received as input and shortage of semantic knowledge in the system. The underlying syntactic parser of ETAP considers internally many syntactic trees permitted by the grammar and then selects one with the highest probability to be the correct one. However, if this final choice goes wrong, the enconverter is stuck with garbage input, unless the user switches to interactive mode and manually corrects the syntactic tree. The system has no means to judge, which of the alternative syntactic trees is the best from the semantic point of view. The other problem of insufficient semantic knowledge affects word sense disambiguation and limits the system's ability to handle complex phenomena, such as light verbs and scopes of modifiers. Only 3.1% of polysemic English words in the dictionary have word sense disambiguation rules. In order to achieve best possible performance the density of such rules must be increased. Some rules cannot be adequately formulated without extra ontological knowledge. The ontology associated with the UNL dictionary is going to provide the necessary information but it is still a work in progress.

Bibliography

Blanc E. 2005. About and around the French enconverter and the French deconverter. *Universal Networking Language: advances in theory and applications* Mexico City: National Polytechnic Institute;157 – 166

Boguslavskij, I.M. 1996. Sfera dejstvija leksicheskih edinic. *Shkola «Jazyki russkoj kul'tury», M.*

Boguslavsky I.M., Dikonov V.G. 2008. Universal Dictionary of Concepts *Proceedings of the first MONDILEX workshop "Lexicographic Tools and Techniques"*. M.; 31 – 42

Dikonov V.G. 2008. UNL Graph Structure *Informacionnye processy* Vol.8 Issue.1; 84 – 97

Dikonov V.G. 2009. Atributy modalnosti v UNL *Informacionnye tehnologii i sistemy* Bekasovo; 230 – 238

Ju.D. Apresjan, I.M. Boguslavskij, L.L. Iomdin, A.V. Lazurskij, L.G. Mitjushin, V.Z. Sannikov, L.L. Cinman, 1992 *Lingvisticheskij processor dlja slozhnyh informacionnyh sistem.* M.: Nauka.

Mel'čuk, I. A. 1974. Ob odnoj lingvisticeskoj modeli tipa "Smysl<=>Tekst" *Izvestija Akademii nauk SSSR. Serija literatury i jazyka*. M. AN SSSR, 1974. — Vol.33. Issue.5; 436 – 447.

UNL Specifications 2005. http://www.undl.org/unlsys/unl/unl2005

Le Morpheur: An Online Tool to Teach French Verbal Inflectional Morphology

Stephanie Doyle Lerat

Dalhousie University
Halifax, Canada
sjdoyle@dal.ca

Abstract

This paper describes a web-based conjugator which was developed using insight from the Meaning-Text Theory. This website, *Le Morpheur*, seeks to present L2 verbal inflectional morphology to adult Anglophone learners of French by adopting a semantic approach. Participants in a study evaluating this pedagogical resource were enthusiastic about the website indicating that continued development of this tool is desirable.

Keywords

Verbal inflectional morphology, Second language acquisition, French L2, Computer-assisted language learning, Meaning-Text Theory

1 Semantic Approach

Language allows a speaker to express meaning. In order to effectively express a meaning the grammatical rules of a given language must be respected. For example, if a speaker wishes to express the future tense in French, there is a verbal suffix which conveys this meaning (along with the grammatical meaning 'indicative'), **-r**, as in, *Jean finira son travail dans trois jours* 'Jean will finish his work in three days'.

This system of correspondences between inflectional meanings (grammemes) and their means of expression is known as verbal inflectional morphology. Verbs are very important in French, as in all languages; therefore it is important that French language learners become proficient users of this system.

Researchers in the field of Second Language Acquisition affirm that a pedagogical approach which focuses on the rule-based nature of language is the most effective for adult learners (Norris & Ortega 2000, Spada & Tomita 2010). For the specific case of verbal inflectional

morphology, this finding is important as morphological means are not the sole way to express these meanings and learners in non rule-based teaching environment may not pick up on verbal inflection by themselves. VanPatten (2002) underlines that learners tend to pay more attention to lexical ways of expressing a meaning than to morphological ones. For example, in the sentence *Jeanne te téléphonera demain* 'Jeanne will call you tomorrow', a L2 learner is more likely to pay attention to the lexical means of expressing the future, *demain* 'tomorrow', than the morphological means, the suffix **-r**[1]. According to VanPatten, learners need a pedagogical approach which explicitly focuses on these morphological means of expression. For us, the link between an inflectional meaning and its expression needs to be clarified for learners. This responds to Davis' (2009) call for the further development of semantic approaches to grammar teaching.

Adult second language learners, contrary to child language learners, possess a lot of language knowledge. They have completely mastered at least one other language[2]. As a rational learner, an adult makes hypotheses about L2 based on past experience with L1 (Hashamdar, 2010). It is important, therefore, not only to consider systems which exist in the target language but also to keep in mind the possible sources of correspondence and divergence between L1 and L2. Pedagogical approaches adopting this perspective (Kupferberg & Olshtain, 1996, Corbeil, 2005, Ghabanchi & Vosooghi, 2006, Paradowski, 2007, Kupferberg, 2009) have been effective in assisting learners to develop proficiency with difficult structures in L2.

The objective of our Master's thesis project (Doyle, 2011), which this paper is based on, was to develop a way to assist adult learners of French with verbal inflectional morphology by focusing on the link between the meanings which are expressed through this system and their means of expression. Our main focus was on applying linguistic theory to pedagogy rather than developing a new theory concerning verbal inflectional morphology. Three steps were followed in order to do so. First, the elements of French verbal inflectional morphology were outlined. Next, a means of presenting this system to adult learners was developed. Finally, a prototype of our pedagogical tool was evaluated by French language learners. The rest of this paper will describe how this objective was met by looking at each of these three steps.

2 Theoretical Notions

The Meaning-Text Theory (Mel'čuk, 1997, Milićević, 2006), with its synthetic approach to language, is well adapted to our objective of helping learners comprehend and put into practice the link between inflectional meanings and their means of expression. The formalisms of the Meaning-Text Theory can help us elucidate the meanings expressed through verbal inflectional morphology. (We are assuming the reader has some knowledge of these formalisms.) A simple example can demonstrate this. Let us consider the simplified Semantic Structure of the following sentence (Figure 1):

---

[1] It is quite possible that learners might in fact notice the suffixes expressing both the future and the person/number agreement such as **-rai**, **-ras**, **-ra**, etc.

[2] The field of L3, L4, etc., acquisition is a rich field of study which merits further consideration; however, for the time being, we have only taken into account the case of an adult learning a second language.

(1) *Jeanne téléphonera au médécin.* 'Jeanne will call the doctor'

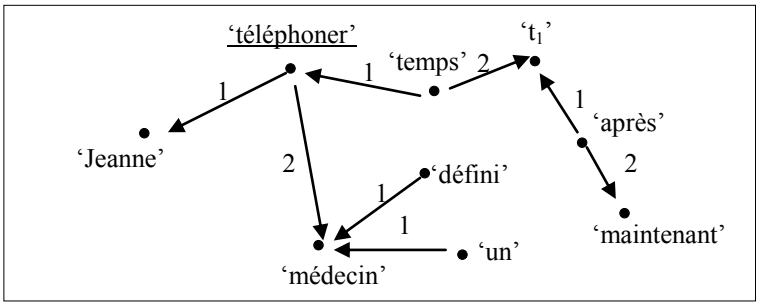


Figure 1 : Semantic Structure of (1)

In this Semantic respresentation, we can see the meaning '[telephoner] après maintenant' ('[to call] after now'). This is an inflectional meaning which will later be expressed through a verbal suffix. Let us now consider the Deep Syntactic Structure of (1):

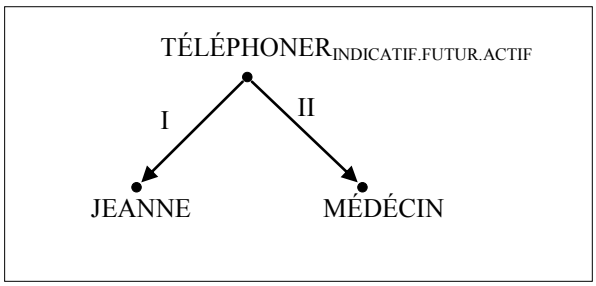


Figure 2 : Deep Syntactic Structure of (1)

In Figure 2, the meaning '[téléphoner] après maintenant' ('[to call] after now') is expressed by the grammeme FUTUR, that is to say, by verbal inflection.

These formal representations allow us to observe the transition of a meaning to its expression; in later levels of representation, it is possible to see each grammeme's means of expression.

A lot of theorisation exists concerning French verbal inflectional morphology. The first big challenge we faced was to determine what was to be considered as part of this system. Our description was based on the theoretical notions outlined in the *Cours de Morphologie Générale* (CMG) (Mel'čuk, 1993-2000). We decided to deal with verbal forms used in Standard French, thus excluding *imparfait du subjonctif* and *plus-que-parfait du subjonctif*. As well, periphrastic expressions such as *être en train de V-infinitif, être sur le point de V-infinitif* and *venir de V-infinitif* were not included. We also did not take into account any of the *surcomposé* forms.

Our preliminary description has been simplified. For a more detailed description of French verbal inflectional morphology based on MTT, see Lareau (2008). Descriptions and discussions concerning this system in other frameworks include, among many others, Vet (1980), Wilmet (1993), and Gosselin (1996, 2005).

A comprehensive description of French verbal inflection morphology requires a clear identification of all its inflectional categories. An inflection category is a set of significations 's' which are expressed with a class **C** of linguistic signs, so that 1) an element of 's' must be expressed with each member of **C** and 2) the elements of 's' are mutually exclusive, i.e., only

one element of 's' can be expressed at a time (CMG I: 262-264). For example, the inflectional category of mood in French contains the following elements: INDICATIF 'indicative', IMPÉRATIF 'imperative', SUBJONCTIF 'subjunctive' and CONDITIONNEL 'conditionnal'. These elements are called *grammemes* and we refer to them using small capitals.

There are two types of inflectional categories: semantic inflectional categories and syntactic inflectional categories. A speaker can freely choose an element from a semantic inflectional category, selecting the element which represents the meaning he wishes to express. Table 1 outlines the four semantic inflectional categories of French verbs as well as their constituent grammemes.

Semantic inflectional categories	**Grammemes**
Mood	INDICATIF, IMPÉRATIF, SUBJONCTIF, CONDITIONNEL
Tense	PRÉSENT, PASSÉ (IMPARFAIT, PASSÉ COMPOSÉ, PASSÉ SIMPLE), FUTUR
Relative Tense	SIMULTANÉ (SIMULTANÉITÉ DANS LE PASSÉ), ANTÉRIEUR (PLUS-QUE-PARFAIT, PASSÉ ANTÉRIEUR, FUTUR ANTÉRIEUR), POSTÉRIEUR (FUTUR DANS LE PASSÉ)
Voice	ACTIF, PASSIF COMPLET, PASSIF SANS AGENT, RÉFLÉCHI DIRECT, RÉFLÉCHI INDIRECT

Table 1 : Semantic Inflectional Categories of the French Verb[3]

An element from a syntactic inflectional category is chosen for grammatical reasons and is not present in the semantic structure of the corresponding sentence. There are four syntactic verbal inflectional categories in French: finiteness, person/number agreement of the finite verb with its subject, gender/number agreement of the past participle with the subject and gender/number agreement of the past participle with the direct object. Table 2 outlines these categories and the grammemes which make them up.

Syntactic Inflectional Categories	**Grammemes**
Finiteness	FINI, INFINITIF, PARTICIPE, GÉRONDIF
Person/Number Agreement of the finite verb with its subject	1.SG, 2.SG, 3.SG, 1.PL, 2.PL, 3.PL

[3] In this table the three grammemes, IMPARFAIT, PASSÉ COMPOSÉ, and PASSÉ SIMPLE, are between brackets because they express a meaning in addition to 'before now'. The grammemes of relative tense between brackets express both a relative tense and a tense. For example, PLUS-QUE-PARFAIT expresses both 'anterior' and 'before now'.

Gender/Number Agreement of the past participle with the subject	MASCULIN, FÉMININ, SINGULIER, PLURIEL
Gender/Number Agreement of the past participle with the direct object	MASCULIN, FÉMININ, SINGULIER, PLURIEL

Table 2: Syntactic Inflectional Categories of the French Verb [4]

The second part of our description involves the grammemes' means of expression, that is to say the formal means used to express these significations. There are four possible means of expression in language: lexical units, prosody (declarative, interrogative, ironic, etc), word order and inflection. The French grammemes above are expressed by two different means: 1) a morphological marker, such as an affix or 2) an analytical form of a lexeme.

An affix is a morph which expresses a grammatical signification. As we have already observed, the French grammeme FUTUR is expressed by the suffix **-r**, (*[il] li+**r**+a* '[he] will read').

An analytical form is a construction which is made up of two or more word-forms. One of these word forms expresses a lexical meaning and the other(s) inflectional meanings. (CMG I: 338). For example, in French, *[j'] ai lu* '[I] read', *[il] aura lu* '[he] will have read', *ayant lu* 'having read', [*elles*] *avaient été lues* '[they] had been read', etc., are all analytical forms of the lexeme LIRE 'to read'. The grammeme PASSÉ COMPOSÉ is expressed through the analytical construction **Auxiliary**$_{\text{INDICATIF.PRÉS}}$ + **V**$_{\text{PARTICIPE.PASSÉ}}$ (*[il] a lu* '[he] read').

In French, some grammemes are expressed cumulatively. This means that there is one morphological means which expresses two or more grammemes. For example, INDICATIF, a grammeme of mood, is expressed cumulatively with the grammeme IMPARFAIT from the inflectional category of tense by the suffixes **-ai** (*[il] lis+**ai**+t* '[he] was reading') and **-i** (*[nous] lis+**i**+ons* '[we] were reading').

3 Le Morpheur: The Website

After having described French verbal inflectional morphology as well as having undertaken a comparative study of this system and English verbal inflectional morphology, we sought to present this information to language learners. As technology is more and more present in Canadian university classrooms, and students respond favourably to computer-based pedagogical tools (Peters *et al.*, 2009), we surveyed existing websites offering to assist language learners with French verbal inflectional morphology. (The results of our study can be found in Doyle, 2011: 30-35.) It became clear that there was no existing website which described this system using a semantic approach.

[4] We have combined syntactic inflectional categories to simplify our description. For example, in the case of the agreement of the finite verb with the subject, the verb agrees with both the number and the person of the subject, these being two separate inflectional categories.

To meet this need, we developed *Le Morpheur*, a web-based pedagogical tool accessible at www.lemorpheur.fr. This section will briefly describe the three main parts of the bilingual website: 1) The semantic-based conjugator, 2) Descriptions of the inflectional categories and 3) Sample exercises.

3.1 Le Morpheur: The Conjugator

The current version of our website's main page, illustrated in Figure 3, is the semantically oriented conjugator we developed.

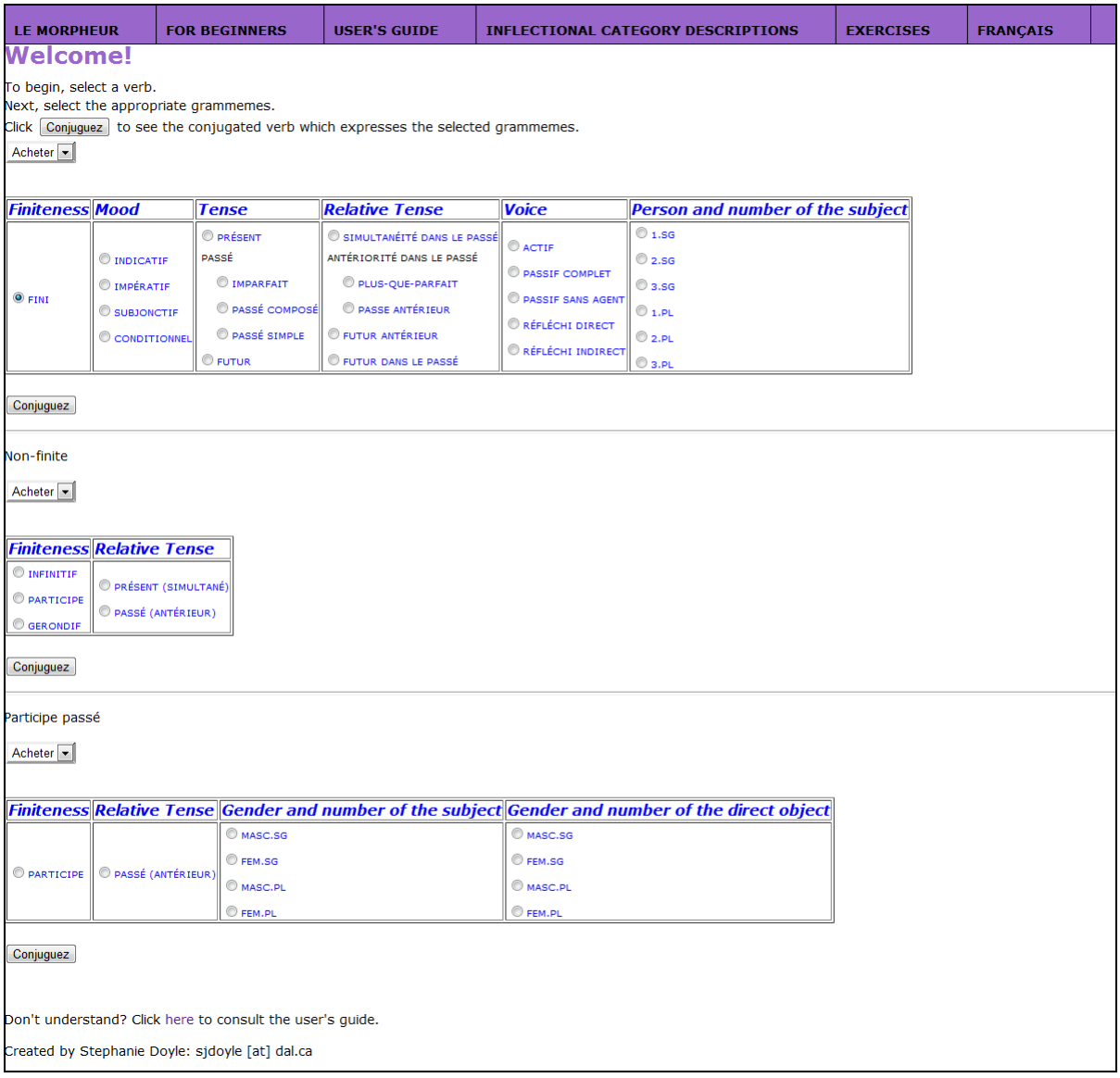


Figure 3 : Le Morpheur

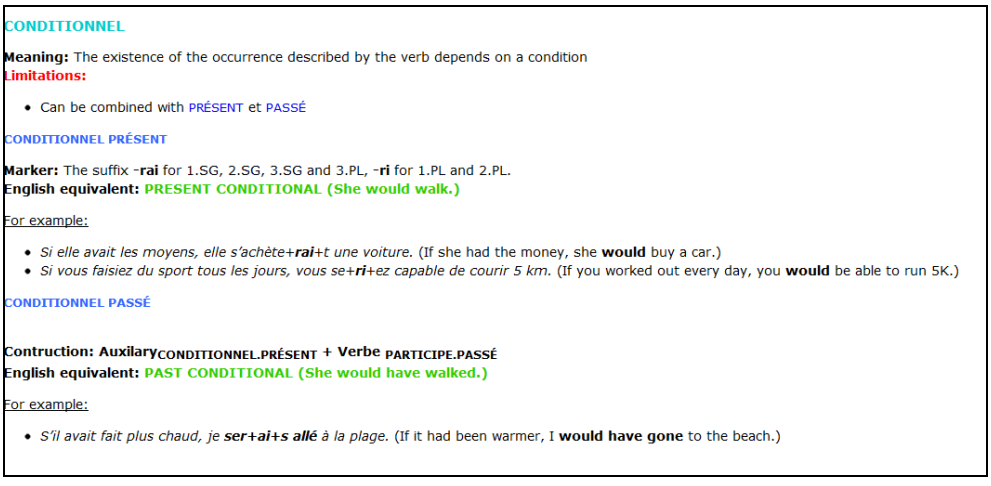
The user is presented with a table of inflectional categories of French and their grammemes. In order to access a verbal form, the user first selects a verb from the drop down menu. The current choice is limited to 10 verbs which represent frequently used verbs from the three conjugation groups that exist in French. Naturally, we intend to expand this choice in the

future. The user next selects the configuration of grammemes representing the meaning he wishes to convey. The user then clicks on "Conjuguez" 'conjugate' and the corresponding verbal form is shown. If the user has a question about the meaning expressed by a given grammeme, there are two options available. When the cursor is passed over the name of a grammeme, it is highlighted and a brief descriptive message appears next to it. For a more in-depth description, the user can click on the name of the grammeme and the description in question will appear in another window.

In order to help learners to understand how grammemes can be combined, the user can potentially select a non-existent configuration of grammemes. In this case, when the user clicks on "Conjuguez", he will see a message explaining the reason why the configuration is erroneous. For example, if the configuration **IMPÉRATIF**.PRÉSENT.ACTIF.**1.SG** (the problem appearing here in bold) is selected, a message will appear reminding the user that the grammeme IMPÉRATIF can only be combined with 2.SG, 1.PL and 2.PL.

3.2 Inflection Category Descriptions

The purpose of the description section is to introduce each inflectional category and describe all the grammemes which make up the category. The layout adopted includes a brief description of the meaning expressed by the grammeme, how it is expressed as well as examples in order to help learners understand how each grammeme fits into the system of verbal inflectional morphology as a whole. As it is important to keep in mind that learners consider L1 in order to facilitate understanding of L2, if there is an English equivalent of a grammeme, it is indicated. If a grammeme does not have an equivalent in L1, this is also indicated. Figure 4 is the description of the grammeme CONDITIONNEL (from the inflectional category of mood).



CONDITIONNEL

Meaning: The existence of the occurrence described by the verb depends on a condition
Limitations:

- Can be combined with PRÉSENT et PASSÉ

CONDITIONNEL PRÉSENT

Marker: The suffix **-rai** for 1.SG, 2.SG, 3.SG and 3.PL, **-ri** for 1.PL and 2.PL.
English equivalent: PRESENT CONDITIONAL (She would walk.)

For example:

- *Si elle avait les moyens, elle s'achète+**rai**+t une voiture.* (If she had the money, she **would** buy a car.)
- *Si vous faisiez du sport tous les jours, vous se+**ri**+ez capable de courir 5 km.* (If you worked out every day, you **would** be able to run 5K.)

CONDITIONNEL PASSÉ

Contruction: Auxilary$_{\text{CONDITIONNEL.PRÉSENT}}$ **+ Verbe** $_{\text{PARTICIPE.PASSÉ}}$
English equivalent: PAST CONDITIONAL (She would have walked.)

For example:

- *S'il avait fait plus chaud, je **ser+ai+s allé** à la plage.* (If it had been warmer, I **would have gone** to the beach.)

Figure 4 : Description of CONDITIONNEL, a grammeme of mood

3.3 Exercises

Two sample exercises based on *Le Morpheur* have been developed to allow users to test their understanding of the way verbal inflectional meanings are expressed in the system presented. The first exercise, "Determining the Verbal Form", models the process of language

production. In this exercise, the user is presented with a meaning, represented by a configuration of grammemes, and must determine the verbal form which expresses this meaning.

The second exercise is called "Determining the Meaning" and reflects the process required in linguistic analysis. In this exercise, the user must, from a given verbal form, determine the expressed meaning, that is to say the configuration of grammemes it expresses.

4 Evaluation by Students

A preliminary evaluation of *Le Morpheur* was carried out with a group of students taking a first-year French course at Dalhousie University (Canada). The purpose of this evaluation was to ascertain language learners' opinion of this resource. Participation was on a voluntary basis. Following a short demonstration of *Le Morpheur*, a two-part questionnaire was distributed to the participants to fill out outside of class and return to the researcher. Part one of the questionnaire was made up of two different types of questions. The first type asked for the verbal form which expressed a configuration of grammemes and the second type for the configuration of grammemes expressed by a given verbal form. Part two of the questionnaire asked the participants to give their opinion of *Le Morpheur*. The participants indicated their level of agreement with a set of five statements from 1 (strongly disagree) to 5 (strongly agree) and were given the opportunity to elaborate upon their reasons for their answer. At the end of the questionnaire, space was provided for additional comments.

The overall reaction to *Le Morpheur* was positive: 10 out of the 13 participants indicated that this tool helped them better understand how French verb conjugation works. The participants were very successful in using and applying the notions presented on the website. Due to the low participation rate and the limited breadth of the questionnaire, it is impossible to draw any definitive conclusions from this evaluation; however, the enthusiasm of the participants is encouraging.

5 Future Work

In this paper, we presented *Le Morpheur*, a pedagogical resource created to help adult Anglophone learners with French verbal inflectional morphology. Our objective is to eventually determine if a semantic based approach to teaching this system is more effective than a traditional approach. Before being able to do so, the site requires further development in the areas of accessibility as well as content.

From an accessibility perspective, the grammeme descriptions could particularly benefit from further work. The development of user-friendly descriptions is challenging because it is difficult to offer descriptions which are both theoretically accurate and accessible to language learners. A remark made by students who evaluated the resource was that the explanations were, by times, too abstract. Future descriptions will be developed with input from language learners in order to ensure their accessibility.

With respect to the content of the site, the description of French verbal inflectional morphology must be developed. The current descriptions are limited in so far as they only take

into consideration one meaning for each grammeme. For instance, the current version of the site describes the grammeme IMPARFAIT, as expressing 'V taking place before now' and the unaccomplished nature of the occurrence, as in *À ce moment-là, Sarah regardait la télé* 'Sarah was watching TV then'. The meaning 'a repetitive occurrence in the past', as in *Sarah regardait la télé tous les soirs* 'Sarah watched TV every evening' has not yet been included on the website. We intend to expand the descriptions to include all possible meanings expressed by each grammeme.

From a pedagogical perspective, it is important to add details concerning the different verbal groups in French which are currently not given consideration on the site. For example, language learners need information concerning changes in verbal radicals.

Following improvements to the site, we wish to undertake a study to determine the effectiveness of *Le Morpheur*. We will compare our semantic approach with the traditional method of teaching French verbal inflectional morphology by measuring the proficiency of two groups of students prior to and following, in one case traditional instruction, and in the other use of *Le Morpheur*.

Two of our goals for the long term are 1) the elaboration of a version of *Le Morpheur* for Francophone adult learners of English and 2) the development of a comprehensive grammar course which integrates this approach.

Acknowledgements

Thank you to Jasmina Milićević, Alexandra Tsedryk and three anonymous reviewers for their insight and helpful comments.

Bibliography

Corbeil, G. 2005. Focus-on-Forms Instruction: Different Outcomes on Constrained- and Free-Production Tasks? *Canadian Journal of Applied Linguistics*, 8(1):27-45.

Davis, J. 2009. Rule and Meaning in the Teaching of Grammar. *Language and Linguistics Compass*, 3(1):199–221.

Doyle, S. 2011. *Pour une approche sémantique de l'enseignement de la morphologie flexionnelle verbale française aux apprenants anglophone adultes*. Unpublished Master's Thesis. Dalhousie University, Canada.

Ghabanchi, Z. & Vosooghi, M. 2006. The Role of Explicit Contrastive Instruction in Learning Difficult L2 Grammatical Forms: A Cross-Linguistic Approach to Language Awareness. *The Reading Matrix*, 6(2):144-153.

Gosselin, L. 1996. *Sémantique de la temporalité en français*. Duculot: Louvain-la-Neuve.

Gosselin, L. 2005. *Temporalité et Modalité*. De Boeck: Bruxelles.

Hashamdar, M. 2010. Rationality and Rational Learner in Second Language Acquisition. *European Journal of Scientific Research*, 41(4):482-489.

Kupferberg, I. 2009. The Cognitive Turn of Contrastive Analysis: Empirical Evidence. *Language Awareness*, 8(3):210-222.

Kupferberg, I. & Olshtain, E. 1996. Explicit Contrastive Instruction Facilitates the Acquisition of Difficult L2 Forms. *Language Awareness*, 5(3-4):149-165.

Lareau, F. 2008. Vers une grammaire d'unification Sens-Texte du français : le temps verbal dans l'interface sémantique-syntaxe. Unpublished Doctoral Dissertation. Université de Montréal, Canada.

Mel'čuk, I. 1993-2000. *Cours de morphologie générale, vol. 1-5*. Montréal: Les Presses de l'Université de Montréal/Paris: CNRS Éditions.

Mel'čuk, I. 1997. *Vers une linguistique Sens-Texte. Leçon inaugurale (given on Friday January 10th 1997)*. Collège de France, Chaire internationale.

Milićević, J. 2006. A Short Guide to the Meaning-Text Linguistic Theory. *Journal of Koralex*, 8:187-233.

Norris, J. & Ortega, L. 2000. Effectiveness of L2 instruction: A Research Synthesis and Quantitative Meta Analysis. *Language Learning*, 50(3):417-528.

Paradowski, M. 2007. *Exploring the L1/L2 Interface. A Study of Polish Advanced EFL Learners*. Unpublished Doctoral Dissertation. University of Warsaw, Poland.

Peters, M., Winberg, A. & Sarma, N. 2009. To Like or Not to Like! Students Perceptions of Technological Activities for Learning French As a Second Language at Five Canadian Universities. *The Canadian Modern Language Review*, 65(5):869-896.

Spada, N. & Tomita, Y. 2010. Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis. *Language Learning*, 60(2):263-308.

Wilmet, M. 1993. *Grammaire critique du Français*. Duculot: Louvain-la-Neuve.

VanPatten, B. 2002. Processing Instruction: An Update. *Language Learning*, 52(4):755-803.

Vet, C. 1980. *Temps, aspects et adverbes de temps en français contemporain*. Genève: Droz.

Collocations: A Challenge in Computer Assisted Language Learning

Gabriela Ferraro (1), Rogelio Nazar (2), Leo Wanner (1, 3)
(1) Department of Information and Communication Technologies
Pompeu Fabra University, C/ Roc Boronat, 138, 08018 Barcelona
(2) Institute for Applied Linguistics, Pompeu Fabra University
(3) Catalan Institution for Research and Advanced Studies (ICREA)
<firstname>.<familyname>@upf.edu

Abstract

The correct use of collocations is one of the most difficult tasks that the student faces when learning a second language, such that one of the goals of Computer Assisted Language Learning (CALL) is to develop programs that aim to identify collocation errors in learners' writings and propose corrections. However, while statistical models currently used by most of these programs still manage to predict, with a reasonable probability, whether a given word combination is a valid collocation in the language in question or not, they fail to suggest corrections. At most, they offer a list of supposedly valid collocations of the base of the erroneous collocation, from which then the learner shall pick one. This is clearly unsatisfactory. We present ongoing work in which we aim to develop algorithms that do better in that they use the sentential context of the erroneous collocation to suggest a correction and in which we assess how crucial the use of Lexical Functions in the sense of the Explanatory Combinatorial Lexicology is in the context of CALL. All our work is tested on a corpus of American English learners of Spanish

Keywords

second language learning, CALL, collocations, lexical functions, Spanish

1 Introduction

Long time, the research in second language learning in general and in Computer Assisted Language Learning (CALL) in particular focused on difficulties of learners with grammatical constructions. The consequence of this was that while for typical grammatical errors more or less detailed analyses have been performed, all types of errors related to the lexicon have been generally classified as "lexical errors", without any further distinction (Granger, 2007). This is certainly a gross oversimplification. One of the larger classes of lexical errors is constituted by errors in the use of collocations (Granger, 1998; Nation, 2001). Since the early 2000ies, a considerable amount of work has been carried out in CALL on the development of programs

(although focused mainly on English as L2[1]) that judge a combination to be a valid or invalid collocation and, in the latter case, attempt to provide a list of correction suggestions. But, again, to consider all collocation errors to be of the same unique class is an oversimplification which does not do justice to the complexity of the problem and thus to the needs of learners. Alonso Ramos et al. (2010) presented a fine-grained collocation error typology which is based on an empirical study of a corpus of American English learners of Spanish (Lozano, 2009).[2] This typology reveals that learners often literally translate collocation elements from their native tongue, use non-existing words as collocation elements, get a wrong subcategorization for one of the elements, etc. Each of these errors requires a potentially distinct focus of the learning aid offered to the learner. Furthermore, in order to be able to correct an error in a targeted way, the meaning that the learner intended to express by the erroneous collocation must be known. In other words, we need to know the Lexical Function (LF) that the learner intended to use. In order to facilitate both learning aids that react to each type of collocation error distinctly and programs that are able to detect and correct collocation errors, the work in the COLOCATE project focuses on the following two tasks: (i) annotation of a learner corpus with collocation error types as defined in (Alonso Ramos et al., 2010) and with the corrections of the errors (tagged additionally with LF labels); (ii) development of algorithms for automatic recognition of collocation errors and their correction – in a long term, at the level of LFs. The first task is addressed by Alonso et al. (Alonso et al., 2011; Vincze et al., 2011). In what follows, we focus on the second task, presenting the state of our current effort towards this long term goal and assessing the next steps to be taken. In the next section, we discuss the related work in the area of collocation checking and correcting. In Section 3, our approach is outlined and its advances compared to the state of the art are discussed. Section 4 finally, presents the lines of our future work in this area.

2 Related Work

The research in the area of collocation checking focused so far mainly on one of the tasks related to collocation error correction: assessment whether a given word combination is a valid collocation in L2. The task of correction has been accounted for, as a rule, cursorily in that a list of collocations of the base in question to choose from has been offered.

The task of the validation of a word combination as a collocation is closely related to the task of collocation identification. Outside CALL, the identification of collocations in corpora has been actively worked on since the late eighties. The majority of the works explore purely statistical models (Choueka, 1988; Church & Hanks, 1990; Evert, 2007; Pecina, 2008). These ("first generation") models can be more or less complex, but all of them measure in one way or the other the distribution of words in combination and in isolation. Some of the works

[1] Following the terminology in language learning, we refer to the native tongue of the learner as L1 and the language being learned as L2.

[2] The corpus is question was CEDEL2 (http://www.uam.es/proyectoinv/woslac/cedel2.htm), which has been compiled by the group directed by Amaya Mendikoetxea from the Universidad Autónoma de Madrid. It contains about 400.000 words of essays in Spanish on a predefined range of topics by native speakers of English.

combine the statistical model with the use of some syntactic features – e.g., submitting to the statistical model only words in collocation-valid syntactic structures (Smadja, 1993; Kilgarriff, 2006; Evert and Kermes, 2003). Most recent statistical proposals take the context of the words that tend to occur into account, which allows for an indirect consideration of the semantics of these words (Bouma, 2010). Another strand uses the co-occurrence range of a given word, i.e., relative frequencies of tokens that co-occur with this word most often (Wible and Tsao, 2010). Opposed to the frequency-based models above is our previous work (Wanner, 2004; Wanner et al., 2005), which uses explicit semantic information from EuroWordnet (Vossen, 1989) to identify and classify collocations with respect to the typology of LFs.

In CALL, the vast majority of the approaches uses statistical models of the first generation (see Chang et al., 2008; Chen, 2010; Park, 2008 and others) or do not use Natural Language Processing techniques at all. Since the pioneering work by Shei and Pain (2000), quite a few proposals have been made on how to improve the collocation competence of the learner of English. First of all, V+N collocations have been considered; see, for instance, (Park et al., 2008; Chang and Chang, 2004; Chang et al., 2008; Chen, 2010, Wu et al., 2003; Wu et al., 2010; Wu, 2010). Futagi et al. (2008) are among the few who treat other syntactic constructions and also consider grammatical errors related to collocations. As far as the resources used in these proposals are concerned, the tendency is to use, in addition to the learner corpus, synonym dictionaries, bilingual dictionaries (which shall facilitate the detection of calques from L1) and reference corpora for L2. S. Wu et al. (2010) and Park et al., (2008) use lists of *n*-grams as reference corpus (in the case of S. Wu et al., provided by Google as mirror of the web).

In general we can state that the current proposals on collocation error recognition and correction still suffer from three shortcomings. First, they are not able to distinguish between "true" collocations and frequent free co-occurrences. Consider, in this context, the reaction of the MUST collocation checker (Wu et al., 2010) to the erroneous collocation *make question* in Figure 1,[3] next page. Second, they are not able to offer any kind of error classification, although such a classification would be helpful to find the most adequate correction of the error. Compare, for instance:

(1) a. *Yo tengo el deseo personal de ser bilingüe*, lit. 'I have the personal wish to be bilingual'
b. *gastar todo el año estudiando español*, lit. 'spend all the year studying Spanish'
c. *hablar un lenguaje*, lit. 'speak (a formal) language'
d. *derechos mujeriles*, lit. 'women rights'

where in (1a), we encounter a register error (the collocation *tener* [*un*] *deseo* 'have [a] wish' exists, but it is not appropriate in this context), in (1b) a collocate error (*pasar* [*un*] *año* instead of **gastar* [*un*] *año*), in (1c) a base error (*hablar* [*una*] *lengua* instead of **hablar* [*un*] *lenguaje*), and in (1d) the use of a non-existent word as collocate (*derechos de las mujeres* instead of *derechos *mujeriles*).

[3] Consulted at http://miscollocation.appspot.com on July, 20th 2011.

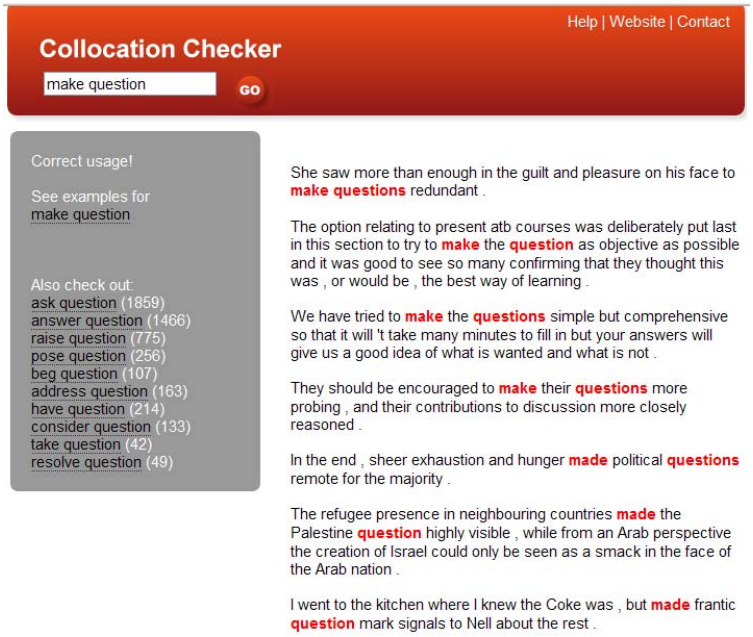


Figure 1: Output of the MUST collocation checker

Third, they are not able to correct an error and must thus stick to offering a list of possible options the learner has to choose from, without any meaningful preferences. Compare, for instance, the following list provided by MUST for the correction of the erroneous collocation *have* [*an*] *obstacle* (in the cited order):

> *overcome obstacle*, *present obstacle*, *clear obstacle*, *jump obstacle*, *prove obstacle*

As further options, the following combinations are given under the heading "Also check out":

> *remove obstacle*, *place obstacle*, *remain obstacle*, *surmount obstacle*, *face obstacle*, *avoid obstacle*, *encounter obstacle*, *eliminate obstacle*, *negotiate obstacle*, *erect obstacle*.

Apparently, MUST attempts to separate correction candidates that are closer to the sought collocation (according to a specific metric) from less likely candidates. But, at least in this example, it does not recognize the intended semantics of the erroneous collocation. The right correction, *face obstacle*, is listed as fifth in the secondary "Also check out"-list. If the level of the learner is not advanced enough, he will not be able to make the right choice.

3 A step forward in collocation error recognition and correction

In our experiments, we focused so far on the third of the three shortcomings of the state-of-the-art proposals listed above, and, since the motivations for the third and second shortcomings are at least related, partially also on the second. Our technique for the recognition of collocation errors is therefore still largely comparable with the state-of-the-art techniques in the field in that it is based on frequency oriented metrics to decide whether a given combination is a correct collocation or not. Similarly to Park et al. (2008) and S. Wu et al. (2010), we use a list of *n*-grams (with $n \leq 4$) as a reference corpus. In our case, this list has been derived from a large Spanish newspaper corpus. Furthermore, we use a number of

auxiliary resources: the Open Office thesaurus, an automatically compiled bilingual Spanish-English dictionary, the Spanish EuroWordNet, and the Web as an additional reference corpus.

In the next subsection, we present first the procedure for the assessment of the correctness of a collocation in Spanish and for the selection of the best correction candidate in case the collocation is judged incorrect, and illustrate then how the procedure performs in action.

3.1 Collocation error and correction procedure

Given a V+N, V+Adv, Adj+N or Adj+Adv combination *C* (extracted from the learner corpus or introduced via an on-line interface) the procedure is as follows:

1. Check whether the relative frequency f_C of C:=Co+B[4] in the *n*-gram list is higher then an empirically determined threshold *T*
2. IF $f_c > T$, *C* is considered a correct collocation of Spanish
   ELSE do
   Collect the synonyms Co_{syn} of *Co* from the auxiliary resources.[5]
   Check whether any $c_{syn} \in Co_{syn}$ forms together with *B* a valid collocation (again, in terms of relative frequency).
   IF there are several valid collocation candidates c_{syn}+B, choose the one judged best according to a number of metrics.

Three different metrics have been applied to judge which of the candidates c_{syn}+B is the best correction of the supposedly erroneous *C*. We present each of them in what follows.

A. Affinity metrics: For each c_{syn}, its affinity is calculated as the product of association strength to *B* and graphic similarity to *Co*, plus the synonymy factor with respect to *Co*. The association strength between c_{syn} and *B* is obtained using the standard *log*-likelihood measure:

$$f(c_{syn} + B) / \ (\text{sqrt}\ (\ f(c_{syn})) * \text{sqrt}\ (f(B))\ )$$

The graphic similarity between c_{syn} and Co is calculated as the Dice coefficient:

$$sim(Co, c_{syn}) = 2\ |Co| \cap |\ c_{syn}| \ / \ |Co| \cup |\ c_{syn}|$$

The synonymy factor of c_{syn} with respect to *Co* is '1' if c_{syn} is among the synonyms of *Co* in the synonym list obtained from the auxiliary resources and '0' otherwise.

B. Lexical context metrics: The lexical context-oriented metrics is grounded in the assumption of distributional semantics, namely that the semantics of a word combination can

---

[4] 'Co' stands for "collocate" and 'B' for "base". In V+N and Adj+N combinations, N is considered the base and V respectively Adj the collocate. In V+Adv, V is the base and Adv the collocate and in Adj+Adv, Adj is the base and Adv the collocate.

[5] In the case of the bilingual Spanish-English dictionary, the "synonyms" of *Co* are the Spanish translations of the English translation equivalents of *Co*. We are aware that this procedure provokes a lot of noise since it ignores the problem of polysemy. However, it has the advantage that it allows us to capture calques from L1.

be approximately deduced from the sentential context in which this word combination appears. Consider, for illustration, the following sentences (from the web) in which one of the words has been removed:

(1) a. *She * a conference on the situation of women rights …*
b. *Mr. White responded to the changing industry and * a conference of critical success.*
c. *Eventcorp** a conference *that met the Conference Committee's criteria.*

(2) a. *The mailman** apples*, bananas, and coconuts.*
b. *Oo baby, here I am, signed, sealed *, I'm yours, oh I'm yours*… [Stevie Wonder song]
c. *Fast Flowers * fresh flowers for every occasion*.

In (1a-c), we can deduce with a certain probability that the missing word is [*to*] *deliver* or any other support verb that goes with *conference*. Why? Distributional semantics suggests that it is the context that allows us to come up with [*to*] *deliver*. In contrast, in (2a-c), this is not the case: we cannot reliably guess the missing verb. This gives us a hint that in (2a-c) the missing verb does not participate in a collocation. We can thus hypothesize that context can be useful for the detection of collocations, or, in our case, for the search of the most adequate correction candidate. More precisely, we assume that given the sentential context c_1, c_2, …, c_n of *Co* in the original sentence of the learner, the candidate c_{syn} with the highest affinity to c_1, c_2, … c_n in a reference corpus is the most adequate correction of *Co* (with "affinity" meaning the highest relative co-occurrence frequency).[6] In contrast to information retrieval oriented search, we do not eliminate from the context the functional words (which are otherwise considered as "stop words" that do not contribute to the quality of the search) since they are essential for our task. For instance, in the learner sentence (3)

(3) *Afortunadamente, su profesora estuvo dispuesta a venderlas y pudo comprar dos máscaras para extender nuestra colección*
lit. 'Fortunately, his professor was willing to sell them and he could buy two masks to extend his collection'

the collocation **extender* [*una*] *colección*, lit. 'extend a collection [of art]' is not correct; this is identified in the first stage of the program. To find the right correction, the contexts of valid collocations of *colección* from our *n*-gram list are examined in the reference corpus with respect to the occurrence of *máscaras*, *para*, and *nuestra* in their neighbourhood. The strongest lexical affinities of *completar* [*una*] *colección*, lit. 'complete a collection' and *ampliar* [*una*] *collection*, lit. 'extend a collection' suggest that the program is accurate in this case.

C. Context feature metrics: As the lexical context metrics, the context feature metrics is based on the idea of distributional semantics. However, in contrast to the lexical context metrics, it allows for a more flexible implementation and the consideration of other features than concrete words. Given the sentential context c_1, c_2, c_n of *Co* in the original sentence of the learner and a list of candidates C_{syn}, the idea is to assess whether any of the contextual

[6] In our preliminary experiments, we used n ≤ 8 (with maximally 2 tokens to the left and 2 tokens to the right of each element of the combination, always within the boarders of a single sentence; duplicates are eliminated).

features $c \in Co$ speaks for the preference of one of the candidates, c_{syn}. For this purpose, we find the maximal probability of each feature c, given a collocation candidate (c_{syni}, b_i). (c_{syni}, b_i) can be calculated as:

$$argmax_{\,i=1,\ldots,n;\, c \in Co}\ (N(c_{syni},b_i) / \Sigma_{j=1,\ldots n} N(c_{synj}\ b_j)) \times (N(c, (c_{syni}, b_i)) / N(c_{syni}, b_i)),$$

where $N(c_{syni}, b_i)$ stands for the number of times the combination (c_{syni}, b_i) occurs in the corpus, and $N(c, (c_{syni}, b_i))$ for the number of times the feature c and the combination (c_{syni}, b_i) co-occur in the corpus at a distance of at most three tokens from each other. For instance, in the learner sentence (4), the collocation **sacarse* [*una*] *operación*, lit. 'take off an operation' is not correct:

(4) *Es fácil, sólo hay que sacarse una operación como Michael Jackson*
lit. 'It is easy, you have only to take off an operation as Michael Jackson'.

To find the right correction, the affinity between the candidate collocations of *operación* 'operation' and each of the contextual features is examined in the reference corpus; e.g.,

[*hay*] … *realizar una operación*, [*que*] *realizar una operación*
[*hay*] … *hacer una operación*, [*que*] *hacer una operación*

(the contextual features are here *hay* and *que*, respectively). The candidate collocation which achieves the highest score is considered to be the correct one.

3.2 Examples of the collocation error correction procedure in action

Let us illustrate the application of the procedure described above to a real world example. (5) is a sentence taken from our learner corpus:

(5) *En mi nueva posición, yo hice planes de viajar para los grupos, acudí el teléfono e hice citas para conferencias con otras compañías para Gary.*
lit. 'In my new position, I made plans to travel for groups, [I] turned to the phone and made appointments for conferences with other companies for Gary'.

One of the potential collocations detected by the program is the V+N combination *hacer citas*, lit. 'make appointments'. Due to its low frequency in the reference corpus, the combination is judged to be a collocation error. In order to find the appropriate correction, all verbal co-occurrences of the base *cita* 'appointment' are retrieved from the reference corpus and filtered; only combinations with synonyms (according to our auxiliary resources) of *hacer* 'make' are kept. The remaining combinations are assessed with respect to their collocation status and non-collocations are removed. The remaining set of combinations includes:

realizar [*una*] *cita* 'realize [an] appointment', *producir* [*una*] *cita* 'produce [an] appointment', *dar* [*una*] *cita* 'give [an] appointment', *tener* [*una*] *cita* 'have [an appointment]', *ir* [*a una*] *cita* 'go [to an] appointment', *acudir* [*a una*] *cita* 'turn [to an] appointment', *declarer* [*una*] *cita* 'declare [an] appointment', *haber* [*una*] *cita* 'receive [an] appointment', *concertar* [*una*]

cita 'arrange [an appointement', *ser* [*una*] *cita* 'be [an] appointment', *agenciar* [*una*] *cita*[7] 'mediate an appointment'.

Given that the remaining set contains more than one option, the best correction candidate is chosen applying the metrics introduced above. The affinity metric suggests *realizar* [*una*] *cita*, while the lexical and context feature metrics suggest *concertar* [*una*] *cita*, which is, in fact, the most appropriate correction of *hacer citas*. Consider a number of further examples summarized in Table 1.

Collocation error	**Suggested correction (collocate)**		
	Affinity metrics	**Lexical metrics**	**Context feature metrics**
realizar meta 'realize goal'	**hacer* 'make'	+*alcanzar* 'reach'	+*alcanzar* 'reach'
cambiar [*al*] cristianismo 'change to Christianity'	+*convertir* 'convert'	+*convertir* 'convert'	+*convertir* 'convert'
comer café 'eat coffee'	+*tomar* 'take'	+*tomar* 'take'	**estar* 'be'
quedar [*la*] *tradición* 'remain [the] tradition'	+*seguir* 'follow'	+*seguir* 'follow'	**pasar* 'pass'
utilizar [*la*] *oportunidad* 'use [the] oportunity'	+*aprovechar* 'take advantage'	**ver* 'see'	**dar* 'give'
concluir [*un*] *problema* 'conclude [a] problem'	+*resolver* 'resolve'	+*solucionar* 'solve'	+*acabar* 'terminate'
empezar [*una*] *familia* 'begin [a] family'	**acomodar* 'accomodate'	+*formar* 'form'	+*formar* 'form'
interrumpir [*una*] *regla* 'interrupt [a] rule'	**establecer* 'establish'	**imponer* 'impose'	+*violar* 'violate'

Table 1: Examples of the correction of collocation errors by our program ('*' stands for wrong correction suggestion and '+' for correct correction suggestion)

Note that some wrong correction suggestions might be valid collocations (as, e.g., *dar* [*una*] *oportunidad*), but with a different semantics than the one required.

3.3 Evaluation

A quantitative evaluation of the procedure for the identification of collocation errors reveals that we are able to judge whether a combination is a correct or incorrect collocation in Spanish with an accuracy of 0.90. Thus, from 61 samples, the procedure fails in six cases. In five of these six cases, correct collocations have been judged to be incorrect. This is mainly due to our purely frequency-based collocation criteria. For instance, *apretar* [*los*] *dientes*, *contar cuentos*, *dar* [*la*] *bienvenida*, and *preparer* [*la*] *comida*, are correct collocations in Spanish, but their frequencies in our reference corpus are too low to consider them valid. On the other hand, for example, *pasar* [*la*] *navidad* is judged by the program to be a correct

[7] The suggestion of **agenciar* [*una*] *cita* as a possible correction candidate is due to the wrong PoS tagging of the bigram *agencia cita* 'agency cites', which is very common in a newspaper corpus as ours.

collocation due to its high frequency in our corpus, although it is questionable in European Spanish.[8]

For the second stage, i.e., the error correction stage, we performed so far two evaluations. First, in order to be able to compare our framework directly with other approaches, we evaluated the accuracy with which we are able to provide lists of valid collocations within which the right correction is encountered. This accuracy amounts to 0.73: in 73% of the trials, the right correction was encountered in the list of possible options offered by our program. Second, we evaluated the capacity of our algorithm to offer the right correction using the context feature metrics, with features being simply words in the original sentence of the learner (the metrics was thus equivalent to the lexical context metrics). The accuracy was 0.542. This is certainly still too low to be used in practical CALL. However, it is to be pointed out that the potential of the contextual features has not been fully explored as yet: the use of concrete words is too restrictive. The experience from statistical NLP (e.g., parsing and generation) teaches us that combinations of morpho-syntactic categories, grammatical functions and words are more promising. We will carry out experiments in this respect in the near future. Furthermore, it needs to be pointed out that this is the first proposal that attempts to suggest the exact correction of a collocation error (rather than to offer a list of suggestions from which the learner has then to choose).

4 Towards an advanced collocation-oriented CALL

In our experiments, we used so far only a limited amount of linguistic information, namely the morpho-syntactic categories of the elements of the combinations. While this information is necessary it is by far not sufficient. Thus, with only this information at hand, we not able to distinguish between *aprovechar* [*de una*] *oportunidad* 'take advantage [of an] opportunity' and *dar* [*una*] *oportunidad* [*a alguien*] 'give [an] opportunity [to so]' – the first being $Real_1$ and the second $CausOper_1$ in terms of LFs. We need to have access to the semantics of collocations! So far, no techniques have been developed that are able to address this challenge without depending on external lexico-semantic resources. On the other hand, the experiments in (Wanner, 2004) demonstrated that even WordNet, as the biggest resource of this kind, is by far not sufficient. This means that the only promising alternative is the use of stochastic techniques based on *distributional semantics* of the collocations in corpora. Our context feature metric is the first try in this direction, but more and additional features need to be exploited to be able to distinguish between the use of $Real_1$, $Oper_1$, $CausOper_1$, etc.

Our future work in the area of CALL will follow three different strands: first, development of techniques for automatic classification of collocation errors according to Alonso Ramos et al.'s typology (2010); second, development of techniques for automatic semantic classification of collocations identified in corpora; and third, amelioration of our techniques for the automatic correction of collocation errors. The learner corpus annotated by LFs and collocation error types by the group LYS at the University of La Coruña (Alonso Ramos et

[8] However, it is a standard collocation in Argentinean Spanish.

al., 2010) and the LF corpus of the Spanish collocation dictionary DICE (Alonso Ramos, 2009) will be essential for all three tasks.

Acknowledgements

Our experiments have been partially run on the Argo cluster of the Department of Communication and Information Technologies, UPF. Many thanks especially to Silvina Re and Iván Jiménez for their help. This work has been supported by the Spanish Ministry of Science and Innovation and the FEDER Funds of the European Commission under the contract number FFI2008-06479-C02-02 in the scope of the project COLOCATE. COLOCATE is a joint effort by the groups LYS, University of La Coruña and TALN, University Pompeu Fabra, Barcelona. We would like to thank the director of the LYS team Margarita Alonso Ramos for the very fruitful collaboration.

Bibliography

Alonso Ramos, M. 2009. Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario? In Cantos Gómez P., Sánchez Pérez, A. (eds.): *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO, pp. 1191–1207.

Alonso Ramos, M., L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira & S. Prieto (2010). Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of LREC 2010*, Malta.

Alonso Ramos, M, L. Wanner, O. Vincze, R. Nazar, G. Ferraro, E. Mosqueira & S. Prieto (2011) Annotation of Collocations in a Learner Corpus for Building a Learning Environment. In *Proceedings of the Learner Corpus Research Conference*, Louvain-la-Neuve.

Bouma, G. (2010): Collocation Extraction beyond the Independence Assumption. In *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, pp. 109–114

Chang, J. S. & Y.C. Chang (2004): Computer Assisted Language Learning Based on Corpora and Natural Language Processing: the experience of project CANDLE. En *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 15–23.

Chang, Y. C., J. S. Chang, H. J. Chen, & H. C. Liou (2008) An automatic collocation writing assistant for Taiwanese EFL learners: a case of corpus-based NLP technology. *Computer Assisted Language Learning,* 21(3):283–299.

Chen, H. H-J. (2010): Developing an English Collocation Retrieval Web Site for ESL Learners, pp. 25–34

Choueka, Y. (1988) Looking for Needles in a Haystack. In *Proceedings of RIAO '88*, pp. 609–623.

Church, K. W., Y P. Hanks (1990): Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Evert, S. (2007) Corpora and Collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, *Corpus Linguistics. An International Handbook*, Berlin: Mouton de Gruyter.

Evert, S. & H. Kermes (2003) Experiments on Candidate Data for Collocation Extraction. In *Companion Volume to the Proceedings of the 10th Conference of the EACL*, pp. 83–86.

Futagi, Y., P. Deane, M. Chodorow & J. Tetreault (2008) A Computational Approach to Detecting Collocation Errors in the Writing of Non-Native Speakers of English. *Computer Assisted Language Learning*, 21(4):353–367.

Granger, S. (2007) Corpus d'apprenants, annotations d'erreurs et ALAO : une synergie prometteuse. Cahiers de lexicologie, 91(2):465–480.

Granger, S. (1998) Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In *Phraseology: Theory, Analysis, and Applications*, A. P. Cowie (ed.), Oxford : Oxford University Press, pp. 145–160.

Kilgarriff, A. (2006): Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress*, Torino, Italy.

Lozano, C. CEDEL2: Corpus Escrito del Español L2. In: Bretones Callejas, C. M. et al. (eds) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería. Almería, pp. 197–212.

Nation, I.S.P. (2001): *Learning Vocabulary in Another Language*, Cambridge: CUP.

Park, 2008 Park, T., E. Lank, P. Poupart & M. Terry (2008): "Is the sky pure today?" AwkChecker: An assistive tool for detecting and correcting errors. In *UIST '08: Proceedings of the 21st annual ACM symposium on User interface software and technology*, New York.

Pecina, P. (2008): A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, pp. 54–57.

Shei, C.C. & H. Pain (2000): An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2): 167–182.

Smadja, F. (1993): Retrieving collocations from text: Xtract. *Comput. Linguistics*, 19(1):143–177.

Vincze, O., M. Alonso Ramos, E. Mosqueira & S. Prieto (2011) Exploiting a Learner Corpus for the Development of a CALL Environment for Learning Spanish Collocations. In *Proceedings of the eLEX 2011*, Bled, Slovenia.

Vossen, P. (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic, Dordrecht.

Wanner, L. (2004) Towards Automatic Fine-Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*, 10(2): 92–143

Wanner, L., B. Bohnet, M. Giereth. 2006. Making sense of collocations. *Computer Speech & Language*, 20(4): 609–624

Wible, D. & N.L Tsao (2010) Stringnet as a Computational Resource for Discovering and Investigating Linguistic Constructions. In *Proceedings of the NAACL-HLT Workshop on Extracting and Using Constructions in Computational Lingistics*, Los Angeles.

Wu, S. (2010) *Supporting collocations learning*. PhD Thesis, University of Waikato, Hamilton, NZ.

Wu et al., 2003 Wu, J. C., K.C. Yeh, T.C. Chuang, W.C. Shei & J. S. Chang (2003) TotalRecall: A bilingual concordance for computer assisted translation and language learning. In *Proceedings of the 41st ACL Conference*, Sapporo.

Wu, J.-C., Y.C. Chang, T. Mitamura & J. S. Chang (2010) Automatic Collocation Suggestion in Academic Writing. In *Proceedings of the ACL Conference*, Uppsala.

The Continuous Expanse of the World and Language[1]

Irina V. Galaktionova

Moscow State University
ig@philol.msu.ru

Abstract

The paper deals with one type of metonymic transfers namely «BORDER → (adjacent) SPACE» which is illustrated with Russian nouns indicating the outermost part of an object: *kraj* 'edge', *storona* 'side', *zad* 'back' and *ploskost'* 'plane'.

Keywords

Semantics, polysemy, metonymy.

1 The spatial aspect of the world and its reflection in word meanings

The spatial aspect of the arrangement of the world is extremely important for man and finds a variety of means of expression in language. For example, in Russian one could mention at the very least the spatial prepositions *pered* 'in front of', *za* 'behind', *okolo* 'near' and the adverbs *daleko* 'far', *szadi* 'behind', *vverx* 'upward' as well as verbs, in the meaning of which the spatial component occupies a central or a more peripheral position (*primykat'* 'to abut': *Stena primykajet k domu* 'The wall abuts the house' vs. *vojti* 'to go into', *vytaščit'* 'to drag out', *perenesti* 'to transfer'). Of course, there are also nouns with indicate a notion of space: *prostranstvo* 'expanse', *mesto* 'place', *(gorodskaja) ploščad'* '(city) square', *komnata* 'room'.

The linguistic reflection of the spatial arrangement of the world in Russian has been very actively studied and from various points of view (see, for example (Arutjunova & Levontina, 2002; Raxilina, 2000; Vsevolodova & Vladimirskij, 2009)). One aspect of the study is the description of various types of transfers on the basis of which derivative meanings of spatial words are formed. These derivatives can have spatial or other meanings.

[1] This research has been financed by a research program of History and Philology Branch of the Russian Academy of Sciences, a grant from the Russian Humanitarian Scientific Foundation (No. 10-04-00273a), and by the Russian President grant for the Support of Leading Scientific Schools No. НШ-4019.2010.6.

A rather large group of Russian nouns related to the notion of space consists of words which name – in their principal or only meaning – parts of physical or territorial objects and indicate the position of this part relative to the surrounding area: *niz* ‘bottom’, *bok* ‘side’, *veršina* ‘summit’, *konec* ‘end’, *kromka* ‘edge’, *opuška* ‘verge’, etc.

According to the position which the named part occupies, nouns may be divided into two groups: the first consists of words which indicate the outermost part of an object, which delimit it and beyond which the object no longer exists: *podnožije (gory)* ‘the foot (of a mountain)’, *verx (kolonny)* ‘the summit (of a column)’, *poverxnost’ (steny)* ‘the surface (of a wall)’, *granica (osveščonnogo prostranstva)* ‘the border (of a lit space)’. The words just mentioned belong to this group. The second, much smaller group includes words which, on the contrary, name the central part of the object, which doesn’t come into contact with the surrounding space: *seredina (komnaty)* ‘the middle (of a room)’, *centr (kruga)* ‘the center (of a circle)’, *serdcevina jabloka* ‘the core (of an apple)’. The following will deal with a few words of the first group.

2 Two types of nouns: discreteness vs. continuity

Many of the nouns which refer to the outermost part of an object embody the naive notion of native speakers that the surrounding world is, to a great extent, discrete: indeed, the word combinations *konec verjovki* ‘the end of a rope’ or *bok arbuza* ‘the side of a watermelon’ do not imply any existence beyond the end or the side of any other objects, even if the rope is tied to something or the side of the watermelon is laying on the ground. Describing the difference between spatial limits and temporal beginning and end points, N. D. Arutjunova comments that «the end or limit of spatial objects does not necessarily imply the direct passage to another object», as opposed to the temporal limit, which, being the end of one time segment, simultaneously indicates the beginning of the next (Arutjunova, 2002: 5 and passim).

It can even be said that most spatial lexemes of such words not only do not imply the passage to another object but, on the contrary, point to the empty space which limits an object that has *verx* ‘top’, *niz* ‘bottom’, *bok* ‘side’, *poverxnost’* ‘surface’, etc.

However, among the names of the outermost parts of an object there are those which, on the contrary, imply the presence of two objects which are in contact. Thus, for example, the word *granica* ‘border’ has two spatial meanings, each of which focuses on the limit of two spatial entities, located one next to the other, which appears in the government patterns: *granica 1.1 meždu morem i sušej* ‘border 1.1 between the sea and dry land’, *granica 1.2 SŠA i Meksiki* ‘border 1.2 between the USA and Mexico’ (lexemes are numbered, hereinafter, as they appear in (Apresjan et al., 2010)). It’s not by chance that objects, which we think of as bordering on empty space, have no borders, e.g. *kraj odejala <kovra, rukava>* ‘the edge of a blanket <rug, sleeve>’, but not **granica odejala <kovra, rukava>* ‘the border of a blanket <rug, sleeve>’. Such meanings may be considered as the linguistic manifestation of the continuity of space, the replacement of one spatial object by another.

It is interesting, however, that words which don’t imply a passage to another object can be used in certain contexts and even have separate derivative meanings which point to the

continuity of the material world. Such usages and meanings arise as a result of metonymic transfers.

3 Metonymy as a way of reflecting continuity of the world

Metonymy of this type, as well as other types of metonymic transfers were studied in (Padučeva, 2000; Padučeva, 2004). E. V. Padučeva calls this type of transfer: «BORDER [line] → SPACE [which is adjacent]» and cites, as an example, the word *krug* ‘circle’ «which means both border – i.e. circumference and all the internal space delimited by it» (Padučeva, 2004: 166); cf. also *v čerte goroda* ‘within the confines of the city’, *v prežnix granicax* ‘within the previous borders’. Moreover, “not only borders can encroach on neighbouring territory but also other discrete spatial objects, cf. *na reke* ‘on the river’, *na dače* ‘at the country home’, *vstretilis’ na lyžne* ‘(they) met at the ski track’, *pojexal na mel’nicu* ‘(he) drove to the mill’, *pošjol na more* ‘(he) went to the sea’ (in the meaning ‘to the sea shore’), etc.” (Ibid., 167). The author comments that this transfer is not completely productive and is limited by specific syntactic parameters, namely, judging by the examples, it occurs in certain positional and directional contexts.

As is known, if, among the meanings of a word there is a relationship of metonymy, especially a productive one, the word usually is not felt to be polysemic (Padučeva, 2000: 250). However, in the dictionary definition, polysemy of such an origin should be indicated.

Let us look at a few of the words that name the outermost parts of an object from the point of view of the presence of this type of transfer in the system of their meanings.

3.1 The main example: KRAJ ‘edge’

Let us begin with the vocable KRAJ ‘edge’. The principal meaning of *kraj 1.1* is defined as being ‘that extremely small part of a physical or territorial object A1 beyond which there are none of its other parts’, which fully reflects the idea of discreteness. The following contexts provide examples of the use of this lexeme: *kraj odejala <prostyni>* ‘the edge of a blanket <sheet>’, *kraj rukava <jubki>* ‘the edge of a sleeve <skirt>’, *kraj lista <gazety>* ‘the edge of a page <newspaper>’, *kraj taburetki <sidenja>* ‘the edge of a stool <a seat>’, *kraj zonta <šljapy>* ‘the edge of an umbrella <hat>’, *kraj stupen’ki* ‘the edge of a step’, *kraj sceny <pomosta, pričala>* ‘the edge of the stage <platform, pier berth>’, *kraj ogoroda <zaroslej>* ‘the edge of a vegetable garden <thicket>’.

As is evident from the definition, one of the taxonomic types of objects that can have a *kraj* are territorial objects of the type *sad* ‘garden’, *zarosli* ‘thicket’, *pole* ‘field’, *les* ‘forest’, *gorod* ‘city’, *ozero* ‘lake’, *luža* ‘pool’, etc. It is obvious that *kraj polja* ‘the edge of the field’, is that part where the field ends: on one side the field exists, on the other side it does not. Language considers this fragment of space as a part of the very field and establishes the relationship between the words *kraj* and *pole* with a genitive (case) construction, one of the basic functions of which is the expression of the relationship part vs. whole. In the context *zasejat’ <vspaxat’> kraj polja* ‘to sow <to plow> the edge of a field’ we are speaking precisely about the edge in this sense. Cf. also *Po vspaxannomu kraju polja […] exal gusenečnyj traktor, voloča kakuju-to složnuju sistemu iz koljos i ryčagov* (V. Dudincev) ‘Along the plowed edge

of the field […] went a caterpillar tractor, dragging some sort of complex system made up of wheels and levers', *Menja perekidyvajut, kak futbol'nyj mjač s kraja na kraj igrovogo polja* (V. Skvorcov) 'They are tossing me about like a soccer ball from one edge of the playing field to the other'.

However, the same construction *kraj polja* is found in contexts of a different type, which refer to the territory directly contiguous with the edge of the field on its exterior side and which is not a part of the field; cf. *[Jegor] uvidel na kraju polja berjozovyj kolok I pošjol k nemu* (V. Šukšin) 'On the edge of the field [Jegor] saw a brich grove and (he) walked toward it'; *Timonin ne stal vyxodit' na otkrytoje mesto, a pošjol opuškoj lesa po kraju polja* (A. Troickij) 'Timonin didn't bother to walk out into the open space, but set out along the verge of the forest, along the edge of the field'; *Ja otpravilsja domoj čerez perelesok, krajem ovsjanogo polja* (P. Aleškovskij) 'I set out for home across the grove of trees, along the edge of the oat field'.

Out of context, word combinations containing *kraj* and names of territorial entities, especially in directional and positional constructions, are ambiguous: indeed, the word combinations *idti krajem lesa* 'to walk along the edge of the forest' can mean the movement of the subject among trees, i.e. through the forest itself and along the forest (possibly along the road which encircles the forest). Thus, *kraj lesa* is a part of the forest itself and a part of the surrounding space which is directly contiguous with it. In the second instance *kraj lesa* is not a part of the forest.

It is understandable why such ambiguity arises specifically in word combinations with names of territorial entities: such an object, by its very nature has a fixed position and is contiguous with other objects of the same type, as opposed to physical objects such as blankets, hats, etc. which do not have permanent «neighbours». Furthermore, not every word combination of *kraj* plus the name of a territorial entity displays a similar ambiguity, for example: *na kraju ostrova* 'on the edge of the island' means only 'on that part of the island which is its edge'. Compare: *žyt' na kraju ostrova* 'to live on the edge of the island', but not **brosit' jakor' na kraju ostrova* 'to cast anchor on the edge of the island'.

It seems inadvisable to present such uses of word combinations of the type *kraj lesa* 'edge of the forest', which do not refer to a part of the very object, as separate meanings. One of the reasons that this should not be done is linked to systemic considerations: it would be strange to attribute to the words *reka* 'river', *ručej* 'creek', *rodnik* 'spring', *more* 'sea', *mel'nica* 'mill', etc. meanings such as territory contiguous with a given object.

However, as it happens, the word *kraj* has a separate meaning formed as a result of metonymic transfer «BORDER (line) → SPACE (which is adjacent)» and which is not indicated in the well-known dictionaries. This is lexeme *kraj 1.2* 'the edge of a cavity or opening A1 in an object and also the part of this object, contiguous with A1': *kraj otverstija <dyrki, treščiny>* 'the edge of an opening <hole, crack>', *kraja ovraga <kanavy, tranšeji, kotlovana>* 'the edge of a ravine <ditch, trench, excavation>', *kraj vodojoma <lužy>* 'the edge of a reservoir <puddle>', *kraj doliny <kotloviny>* 'the edge of a valley <basin>', *dopolzti do kraja polynji <voronki>* 'to crawl up to the edge of the unfrozen patch of water <crater>'.

From heir definition of this lexeme it follows that, for example, *kraj otversyija* ‘the edge of an opening’, being that part of very opening beyond which there are no other of its parts, means, in addition, the part of the object contiguous with it, into which the opening was made. Thus, *kraj otversyija* turns out to be not only a part of the opening but also of another object, which is not expressed in the genitive case. Cf. *Pulja prošla pod samym donyškom cylindra, probiv jego naskvoz’. Kraja oboix otverstij, prožžjonnyx raskaljonnym svincom, byli buro-žjoltymi* (L. Juzefovič) ‘The bullet passed right under the bottom of the cylinder, breaking right through it. The edges of both holes, burned through with hot lead, were brownish yellow’ (**Kraja cylindra byli buro-žjoltymi* ‘The edges of the cylinder were brownish yellow’ does not have the same meaning); *Sergej podprygnul, ucepilsja rukami za kraj otverstija, podtjanulsja i vybralsja na čerdak* (A. Mel’nik) ‘Sergej jumped up, caught hold of the edge of the opening with his hands, pulled himself up and climbed into the attic’.

In most cases these two meanings are well contrasted: a ravine, abyss, precipice, pit, hole, etc. recesses or openings, by the nature of their «anatomy» may only have *kraj 1.2*. However, there are objects which are capable of having *kraj 1.1* as well as *kraj 1.2*. These are cavities filled with water, e.g. a puddle, pond, lake, cf. *Na protivopoložnom kraju ozera nad vodoj jedva vozvyšalas’ […] central’naja čazt’ zemljanki* (V. Obručev) ‘At the opposite edge of the lake, above the water, the central part of the adobe cottage barely appeared’ (reference is made to ‘the part of a lake contiguous with the shore’, i.e. *kraj 1.1*) and *stojat’ na kraju ozera* ‘to stand on the edge of the lake’ (that is ‘to stand on the shore of the lake’ – *kraj 1.2*); at the same time, the sentence *Kraja ozera zarosli kamyšom <osokoj, travoj>* ‘The edges of the lake were overgrown with reeds <sedge, grass>’ requires, for its interpretation, extra-linguistic meanings about where precisely the reeds, the sedge or any other grass, grew – in the water or near it.

In this way, the meaning expressed by the word combination *kraj X-a* ‘the edge of X’ is determined to a great extend by the taxonomic class of X, that is, ulitmately how necessary the other object «on the other side» of *kraj X-a* is and also, if that object is the focus of attention.

The Russian word *kraj*, in the relationship under consideration, can be compared to the English *verge*, which is used in the contexts *the verge of the sea <of the sky>*, *on the very verge of the roof* (≈ *kraj 1.1*) and in contexts such as *the verge of the precipice* (≈ *kraj 1.2*). Moreover, *kraj 1.2* has an English equivalent, *brink* which «refers to a thin strip, part of the surface, which includes the break-off line just in front of the precipice, the steep slope: *the brink of a precipice*» (Apresjan et a., 1979:74)[2].

[2] Only the two meanings of the word *kraj* which are connected with the metonymic relationship that interest us are being considered here. Of course, these are not all of the meanings connected with the indication of a part of a physical object. Accordingly, not all English equivalents are presented here. For more details, see (Apresjan et al., 2010; Apresjan et al., 1979).

3.2 Some other examples: STORONA ‘side’, ZAD ‘back’, PLOSKOST’ ‘plane’

I would like to present a few more examples of the type of metonymic transfer discussed here.

The vocable STORONA ‘side’ contains two lexemes which combine principally with names of elongated objects. *Storona 3.2* is defined as follows: ‘each of the two parts of object A1, located on the right and left of a hypothetical line or surface which divides this object into symmetrical parts, (and) which differs from the second part by feature A2’. This is illustrated by the word combinations *tenevaja <čjotnaja, protivopoložnaja> storona ulicy* ‘the shady <even (numbered), opposite> side of the street’, *pravaja storona lestnicy <eskalatora, koridora, perrona, proseki>* ‘the right side of the staircase <escalator, corridor, platform, glade>’, *levaja storona tela <grudi, lica>* ‘the left side of the body <breast, face>’. *Storona 3.3*, metonymically associated with *3.2* means ‘each of the two parcels of territory, which are divided by object A1, usually elongated which differs from the second parcel by feature A2’: *s toj <drugoi, protivopoložnoj> storony reki <ozera, proliva,uščelja>* ‘at that <other, opposite> side of the river <lake, strait, gorge>’, *perepravitsja na druguju storonu Dnepra* ‘cross to the other side of the Dnieper’, *Na nemeckoj storone granicy <fronta> stojala tišina* ‘On the German side of the border <front> there was silence’. *Storona 3.3 X-a* ‘the side of X’, as opposed to *storona 3.2*, is not a part of X itself, but marks the territory which is contiguous with X.

Besides lexemes designating a part of the very object, the vocable ZAD ‘back’ also contains the colloquial lexeme *zady 3*, which is used only in the plural and which has the meaning ‘a part of the territorial object A1 or the territory which is contiguous with A1, often not very well maintained, located on the other side of A1, which is opposite the main facade and, for that reason, not usually visible by an observer’. From the definition it follows that *zad 3 X-a* ‘the back of X’ may be not only a part of X itself, as, for example in the word combination: *na zadax kladbišča <ogorodov, bol’ničnoj territoriji, učastka>* ‘in the back of the cemetery <the vegetable gardens, the hospital grounds, the plot>’ but also a parcel of territory contiguous with X, as in the contexts: *na zadax nekazistyx stroenij <zdanij>* ‘at the back of the unsightly edifices <buildings>’; *Svalka vyxodila na zady kanceljariji byvšego ministerstva oxrany korony* (A. & B. Strugeckije) ‘The dump faced the back of the office of the former Royal Ministry of the Guard’. In some instances it is difficult to say if a part of X itself is being referred to, or an area contiguous with it. Thus, the word combination *zady vokzala <rynka>* ‘the rear of the station <market>’ may denote the area behind the building where the station or market is located, as well as the «less elegant» part of the area occupied by the station or the market.

The vocable PLOSKOST’ ‘plane’ contains the lexeme *ploskost’ 1.2* ‘a hypothetical flat surface, which is considered to be an extension of object A1’ and which is found in contexts such as *ploskost’ orbity <ekliptiki, meridiana>* ‘plane of sphere <ecliptic, a meridian>’, *ploskost’ osnovanija (piramidy)* ‘the plane of the base (of a pyramid)’, *ploskost’ vraščenija (kolesa)* ‘the plane of rotation (of a wheel)’. In other words, *ploskost’ 1.2* is not part of any object, rather, on the contrary, the object is a part of it. Moreover, the main lexeme *ploskost’ 1.1* denotes the flat surface of an object (*ploskost’ bumažnogo lista <xolsta, steny>* ‘the flat surface of a piece of paper <canvas, wall>’).

4 Continuity and discreteness: the language and the world described by it

It is well known that the semantic expanse of language is being structured continuously[3], a dictionary description is traditionally based on the notion of the relative discreteness of meanings, although it should have the means necessary to establish the links among various meanings and various types of non prototypic usages which appear within the framework of one or another meaning (this is discussed in the lexicographical concepts of Ju. D. Apresjan, for example in (Apresjan, 2009)). Not only is language itself a continuum, but – in a certain respect – it is the linguistic explanation of the space around us.

Acknowledgements

All the above results were prepared by me while working as one of the authors of the Active dictionary of the Russian language. These results were also discussed during meetings at the Sector of Theoretical Semantics of Institute of Russian Language, Russian Academy of Sciences. I want to express my deepest gratitude to all the participants of these discussions and especially to our project supervisor Ju. D. Apresjan. I also would like to thank Mary Anne Cosentini for her help on the English version of this paper.

Bibliography

Apresjan, Ju. D. 2009. *Issledovanija po semantike i leksikografiji*. V. 1. *Paradigmatika*. Moscow.

Apresjan, Ju. D. et al. 1979. *Anglo-russkij synonimičeskij slovar'*. Moscow.

Apresjan, Ju. D. et al. 2010. *Prospekt Aktivnogo slovarja russkogo jazyka*. Moscow.

Arutjunova, N. D. 2002. Vstuplenije. V celom o celom. Vremja i prostranstvo v konceptualizaciji dejstvitel'nosti. In *Logičeskij analiz jazyka. Semantika načala i konca*. Moscow.

Arutjunova, N. D. & I. B. Levontina (ed.) 2000. *Logičeskij analiz jazyka. Jazyki prostranstv*. Moscow.

Vsevolodova, M. V. & Je. Ju. Vladimirskij. 2009. *Sposoby vyraženija prostranstvennyx otnošenij v sovremennom russkom jazyke*. 3 ed. Moscow.

Padučeva, Je. V. 2000. Prostranstvo v obličiji vremeni i naoborot (k tipologiji metonimičeskix perenosov). In *Logičeskij analiz jazyka. Jazyki prostranstv*. Moscow.

[3] However, «together with continuity in certain areas of the semantic expanse substantial «holes» in other of its areas can be seen» (Apresjan, 2009: 422).

Padučeva, Je. V. 2004. *Dinamičeskije modeli v semantike leksiki*. Moscow.

Raxilina, Je. V. 2000. *Kognitivnyj analyz predmetnyx imen: semantika i sočetajemost'*. Moscow.

Transformational Grammarians and other Paradoxes

Thomas Groß

Aichi University 1
Machihata-cho 1-1, Toyohashi-shi, Aichi-ken, Japan
tmgross@vega.aichi-u.ac.jp

Abstract

This paper argues that morphs can/should be granted node status in tree structures. Most theories of morphology do not do this. For instance, word-based morphologies (Anderson 1992 and others) see inflectional affixes appearing post-syntactically, producing a specific word form based on paradigmatic rules. On the other hand, derivational affixes attach prior to syntax. So-called "bracketing paradoxes" (Williams 1981, Pesetsky 1985, Spencer 1988, Sproat 1988, Beard 1991, Stump 1991) such as *transformational grammarian* concern primarily derivational affixes (here: *-ian*). If a theory can avoid bracketing (or structural) paradoxes by maintaining an entirely surface-based account, then this theory should be preferred over competing theories that posit different levels or strata in order to explain the same phenomenon. This contribution demonstrates that such a surface-based account is possible if the *catena* is acknowledged. The catena is a novel unit of syntax and morphosyntax that exists solely in the vertical dominance dimension.

Keywords

Bracketing paradox, catena, morph catena

1 Introduction

(Williams 1981:219f) is credited with introducing "bracketing paradoxes" to theoretical linguistics. He puts forth examples such as the following:

(1) a. hydroelectricity (2) a. Gödel numbering (3) a. atomic scientist

For example (1a) Williams posits the next structure:

(1) b. [hydro-[electric-ity]]

The problem with (1b), Williams realizes, is that the structure cannot provide the adjective *hydroelectric* because the prefix and the root do not appear within a bracket that excludes the suffix. In order to accommodate *hydroelectric*, (1b) must be rearranged thus:

(1) c. [[hydro-electric]-ity]]

But (1c) would see the stem affix *hydro-* attaching before the root affix *-ity*. Hence assumptions about affix order and semantic composition generate a conflict between structures expressing important properties, which can only be expressed in mutual exclusion, never in combination.

Williams realizes that the same conflict can occur within compounds. In (2a) the stem suffix *-ing* appears inside the second compound part, even though the compound as such is *Gödel number*. Conversely, in (3a) the root suffix *-ist* must appear before the compound *atomic science* is created. Note first that (Williams 1981:219f) assumes that (2a, 3a) have identical structure:

(2) b. [Gödel [number-ing]] (3) b. [atomic [scient-ist]]

The problem with these (b)-structures is that they do not represent the semantic relationships accurately. The bracketing should, rather, look like this:

(2) c. [[Gödel number]-ing] (3) c. [[atomic scient]-ist]

Preferring the (b)-structures over the (c)-structures necessitates disregarding semantic composition, while choosing the (c)-structures over the (b)-structures incurs conflict with the assumption of Level Ordering (the assumed order in which root and stem affix attach). (Spencer 1988:673) concludes that Level Ordering is not implicated in these cases.[1]

The discussion of bracketing paradoxes in the literature starts with (Williams 1981) and extends through (Pesetsky 1985), (Sproat 1988), (Spencer 1988), (Beard 1991), (Stump 1991), (Becker 1993) to recent accounts, for instance in Distributed Morphology (Noyer and Embick 2001) or in HPSG on particle-verb-constructions (Müller 2003). These accounts all acknowledge similar problems with the data. But in its contemporary context, this discussion is a side-show to the problem of whether words are structured in a fashion similar to sentences. The question is whether syntactic principles, in particular headedness, apply to a sufficient degree to morphology. This aspect of Williams' (1981) ideas appears in the "head-debate" between (Zwicky 1985) and (Hudson 1987). This debate is summarized well by (Bauer 1990).

Two trajectories can be distinguished in the accounts just mentioned: the generative camp sees morphology as not much different from syntax and as a result, it assumes that headedness also operates in morphology. (Di Sciullo and Williams 1987), (Di Sciullo 2005), and (Williams 2011) pursue this sort of approach. Williams has become increasingly critical of central assumptions in the generative model, however. The Distributional Morphologists (Halle and Marantz 1993, Embick and Noyer 2001/2007, Harley and Noyer 2003, Embick 2003) are the ones who continue to uphold the central tenets of the "syntactic" approach.

The second camp is known as the Paradigmatic (or Word and Paradigm (WP)) approach; it originates with (Robins 1959) and is best represented by (Matthews 1972),

[1] I disregard Pesetsky's (1985) QF-analysis, and Sproat's (1988) Mapping Principle. The critical discussion of these proposals by Spencer (1988:664-72) strikes me as thorough and accurate.

(Anderson 1992), and (Stump 2001). The WP approach sees inflectional morphology as rule-based (or realizational): the word form of a verb is created by a specific rule (e.g. V+/PAST/ creates *saw*, *made*, and *hinted* equally reliably, whereby more specific realizations of that rule block the more general, i.e. regular, realizations). In order to arrive at an unambiguous treatment of bracketing paradoxes, derivation needs to be addressed. (Stump 1991: 720), for instance, argues that the expression *atomic scientist* is simply the value of the paradigm function of the suffix based on the "morpholexical rule" of the suffix. The problem is, though, that attributive modifiers, such as *tall* in *tall scientist*, will not conform to the same paradigm function.

The discussion now proceeds to a brief, and hopefully accurate, description of the treatment of "person noun" paradoxes (Spencer 1988:673) of the sort illustrated in (1-3) in Meaning-Text Theory. Since an anonymous reviewer requested examples, a third section is devoted to syntactic catenae. Thereafter, a brief outline of catena-based morphology follows, which basically argues that morphs should indeed receive node status on the surface. Once granted node status, the paradoxical aspect of data like (1-3) disappears entirely.

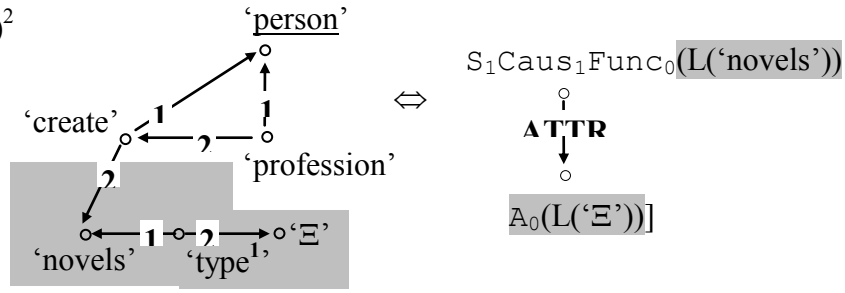
2 Personal nouns in MTT

In Meaning-Text Theory (Mel'čuk 1988/2003, Kahane 2003), bracketing paradoxes are absent due to the multistratal system. According to (Mel'čuk 2009: 2), MTT posits three modules (semantics, syntax, and morphology+phonology), which correspond in the following manner:

(4){$SemR_i$} ⇔ {$SyntR_k$} ⇔ {$MorphR_l$} ⇔ {$PhonR_i$}

The syntax and morphology levels are further divided into *deep* and *surface* structures. For the expression *historical novelist*, which would traditionally be considered a bracketing paradox, MTT posits the next semantic rule:

(5)[2]



The shadowed areas show the lexemes that appear in the resulting expression. The semantic rule on the left shows that the attribute *historical* modifies *novels*. The non-shadowed components on the left show the logical relationship for a person whose profession it is to create *historical novels*. The right side shows the DSyntS, where the attribute depends on the

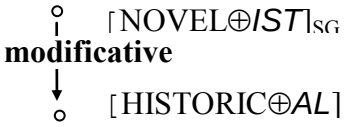
[2] Structure (5) stems from an unpublished manuscript which Igor Mel'čuk kindly provided. Needless to say, anything I say here about this issue reflects my own – perhaps mistaken – understanding of this matter, not Igor Mel'čuk's.

nominal *novels* indexed by the complex LF $S_1Caus_1Func_0$, meaning 'who causes that L begins to exist'.

The predicate 'create' can be substituted against others in order to ensure that *transformational grammarian*, *baroque flautist*, etc. fit rule (5). A transformational grammarian, namely, does not create transformational grammars, but works with them, and a baroque flautist does not make baroque flutes, but plays them. These predicates can be represented by a metavariable. These details are, however, not important.

The DSynt-structure to the right of (5) can then be mapped to a surface structure (SSyntS):

(6)



The italicized suffixes are not yet realized at this level, but rather they stand for groups of suffixes with similar functions; for instance, *IST* may also stand for *-ian*, *-er*, or *-or*, and *AL* can also stand for *-ic*. The Deep Morphological Structure (DMorphS) would therefore look like this:

(7)[HISTORIC⊕*AL*] [NOVEL⊕*IST*]$_{SG}$

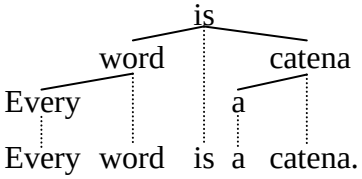
At the DMorphS, the suffixes still function as variables. It is not until the morphological surface structure (SMorphS) that *historical novelist* is realized concretely.

3 Catenae in syntax

This section introduces the notion of the syntactic catena, and it provides evidence that it is a highly salient unit of syntax. Brief examples are provided that show how catenae operate in idiom formation, ellipsis, predicate expansion, and constructions. A catena-based analysis of displacement is already discussed in detail in (Groß and Osborne 2009).

First, however, the concept of the catena is introduced: A catena is ANY WORD OR COMBINATION OF WORDS THAT IS CONTINUOUS WITH RESPECT TO DOMINANCE. This means that any word combination the words of which are connected by immediate dependency relationships qualifies as a catena. Put differently, any tree or subtree of a tree qualifies as a catena. The next example shows how this works:

(8)



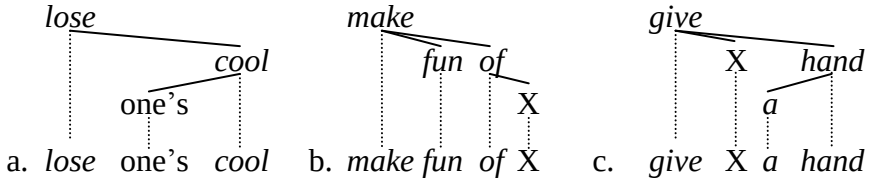
Example (8) contains 15 distinct catenae, all of which are listed here: *every*, *word*, *is*, *a*, *catena*, *every word*, *word is*, *is..,catena*, *a catena*, *every word is*, *word is…catena*, *is a catena*, *every word is…catena*, *word is a catena*, and *every word is a catena*. Every word combination qualifying as a catena constitutes a subtree of continuous, i.e. uninterrupted, dependency

relationships. There are 16 distinct non-catena word combinations in (8), e.g. *every…is*, *every…a*, *every…catena*, *word…a*, *word…catena*, *is a*, *every word…a*, etc. These word combinations fail to qualify as catenae because they are NOT continuous in the vertical dimension, i.e. they do NOT form subtrees.

The following data demonstrate the potential of the catena concept for theories of syntax. Data from idiom formation, ellipsis, predicate expansion, constructions, and displacement are briefly considered; the conclusion will be not only that catenae are singularly important to describe and explain these data, but that the notion of dominance applied in (8) is the accurate one. The points made and data examined are discussed in much greater detail in (O'Grady 1998), (Osborne 2005), (Groß 2010), (Osborne in press), and (Osborne et al. in press.a, in press.b).

The first piece of evidence concerns idiom formation. (O'Grady 1998) shows that many idioms cannot be stored as constituents in the lexicon, but rather they are "chains" (=catenae). The next examples are taken from (Osborne et.al. in press.a):

(9)

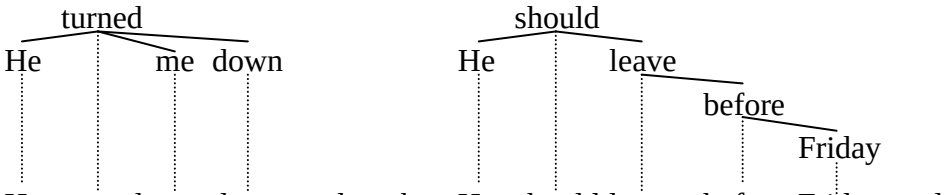


a. *lose* one's *cool* b. *make fun of* X c. *give* X *a hand*

The idioms proper, which are italicized in (9), fail to form constituents: *lose…cool* in (9a), *make fun of* in (9b), and *give…a hand* in (9c) do not qualify as constituents because they fail to include *one's* in (11a), and *X* in (9bc). The idioms proper do, however, qualify as catenae.

Ellipsis is characterized by material missing from utterances. Linguists distinguish many different forms of ellipsis such as gapping, VP-deletion, pseudogapping, stripping, etc. What unifies these ellipsis mechanisms is the requirement that the elided material must form a catena. The next examples are again taken from (Osborne et.al. in press.a):

(10)



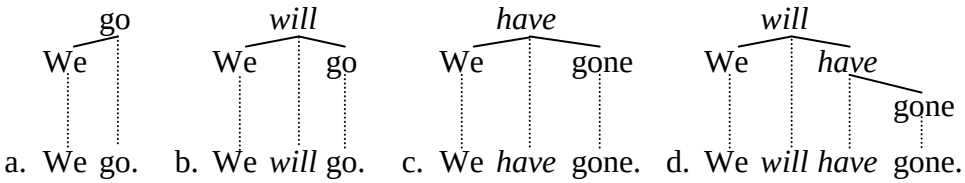
a. He turned me down, and
she … you …

b. He should leave before Friday, and
she should … on Friday.

Example (10a) shows gapping. Note that the elided material in the second conjunct is non-contiguous, hence it fails to form a constituent. Example (10b) shows pseudogapping.

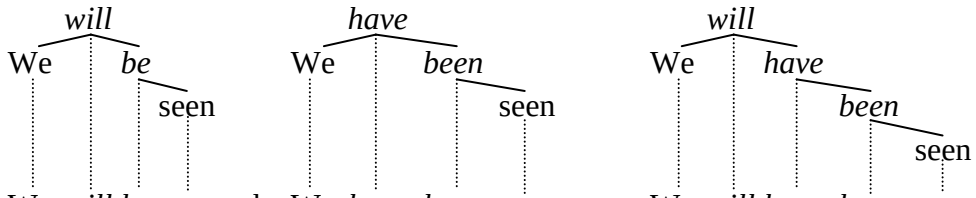
Verbs can be modified in order to accommodate valence, voice, aspect, modality, tense, and/or mood. In English many of these predicate expansions appear as individual words. The verb and these expansions always constitute a catena:

(11)



a. We go. b. We *will* go. c. We *have* gone. d. We *will have* gone.

(12)

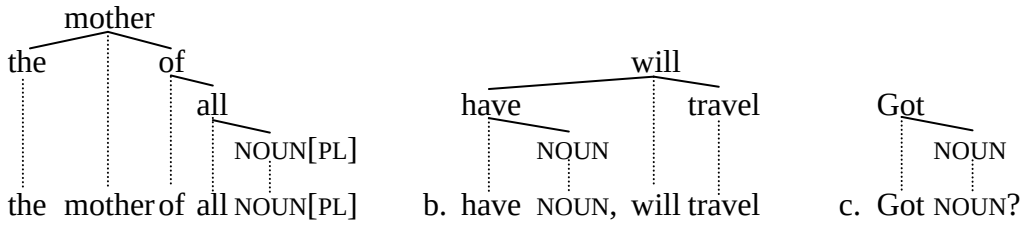


a. We *will be* seen. b. We *have been* seen. c. We *will have been* seen.

The predicate of the verb *go* is expanded by mood/tense (11b), aspect/tense (11c), and mood/tense/aspect (11d). In (12a-c), the passive expands the verb *see*; the passive predicate is expanded by mood/tense (12a), aspect/tense (12b), and mood/aspect/tense in (12c). Note that the expansions (in italics) invariably form catenae, and that they also form catenae together with the verb.

The italicized words in (11b-d, 12) are recognized as constructions in Construction Grammars. Constructions, like idioms, elided material, and predicate expansions, qualify as catenae. A comparatively new construction super-type are *snowclones*.[3] The term was suggested by Glen Whitman in response to a request by (Geoffrey Pullum 2003). Snowclones are phrasal templates that convey clichés by referencing shared cultural knowledge. One famous snowclone is Shakespeare's *To be or not to be*, where any VP can now appear instead of *be*. Other examples include *the mother of all X*, originating from the 1991 Gulf War as *the mother of all wars*, *have X, will travel* from Robert Heinlein's novel *Have spacesuit, will travel*, *Got X?* from the advertisement *Got milk?*. Note that none of the snowclones above qualifies as a constituent; they do, however, qualify as catenae as the next trees show:

(13)



a. the mother of all NOUN[PL] b. have NOUN, will travel c. Got NOUN?

Due to limited space, the above examples must suffice as representative of many other constructions. For a catena-based analysis of constructions, the reader should (Groß and Osborne to appear).

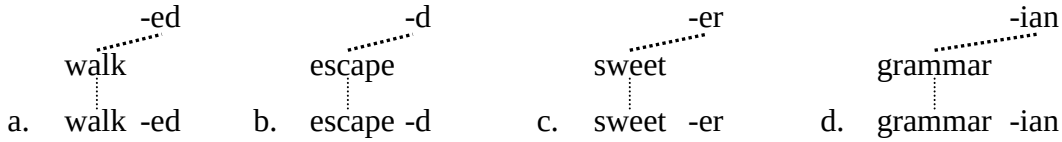
4 Catenae in morphology

The data examined so far suggest that one should explore the possibility of catena in morphology. Are parts of words organized in a fashion similar to words in syntax, i.e. in terms of catenae. And, indeed, it is possible to view the internal structure of complex words in a

[3] http://en.wikipedia.org/wiki/Snowclone

similar fashion. If one replaces "word" with "morph" in the definition of the catena above, one gains the *morph catena*. Two morphs form a morph catena, when the one morph immediately dominates the other. If both morphs are contained within the same word, then *distribution* decides which morph dominates the other. Consider the following examples, where the dotted edges symbolize an *intra-word* dependencies:

(14)

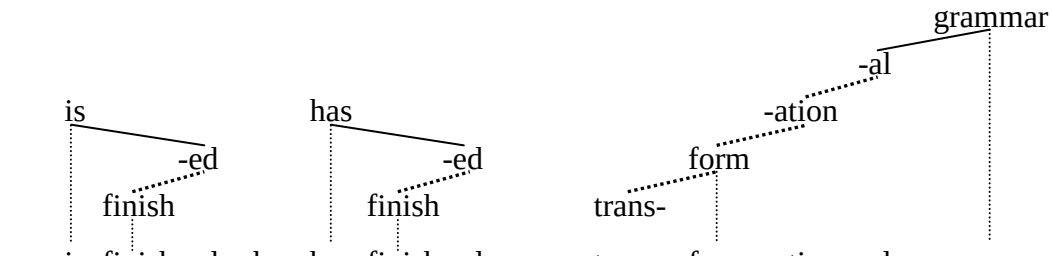


The morph combination *walk-ed* in (14a) distributes like a past tense verb or a participle. In fact, the entire expression distributes like a word marked with *-ed*, rather than as a word in which the morph *walk* appears. The morph tree (14a) further represents the correct semantic scope: the morph *walk* is in the scope of the suffix *-ed*, because the entire expression means the past tense or participle form of *walk*, rather than that a past tense or participle form is engaged in the activity of walking.

The same sort of observation is true for (14b-d). The adjective in (14c) distributes like a comparative adjective, rather than as a positive adjective. The adjective *sweet* is in the semantic scope of the comparative; the entire expression does not mean that the comparative is sweet. The noun in (14d) distributes like a personal noun, rather than as the lexical noun which forms the base. The morph *grammar* is in the semantic scope of the personal suffix; the entire expression does not mean that a person is a grammar. The basis for determining intra-word dominance is thus similar to (Mel'čuk's 2003: 200f) criterion of "surface syntactic dominance" for determining inter-word dominance.

The next case concerns morphs being contained in two different words. Two morphs contained in two different words still form a morph catena if the one morph licenses the appearance of the entire word that contains the other morph. In a sense, this definition builds on (Mel'čuk's 2003: 205) criterion of omissibility and coocurrence. The next examples illustrate *inter-word* morphological dependencies:

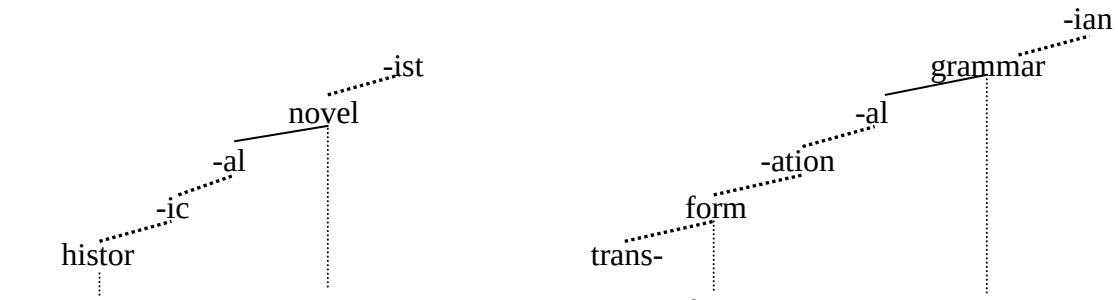
(15)



The morphological structure of *finish-ed* in (15a,b) is an intra-word dependency, and it thus follows the remarks made concerning (14). The adjective *trans-form-ation-al* exhibits three intra-word dependencies. The crucial observation concerning (20c) is that the central morph catenae (that one might expect) are all present: *form*, *transform*, *formation*, *transformation*, *formational*, and *transformational* are all present as catenae. The suffix *-ed* constitutes the root in *finish-ed*, and the suffix *-al* is the root of *trans-form-ation-al*. These suffixes are directly dominated by the morphs *is*, *has*, and *grammar* because the latter license the appearance of the entire words *finished* and *transformational*, of which the suffixes *-ed* and *-al* form the roots.

Personal noun constructions such as *historical novelist*, *transformational grammarian*, etc. cease to be paradoxical on this analysis. Their entire structure is given below:

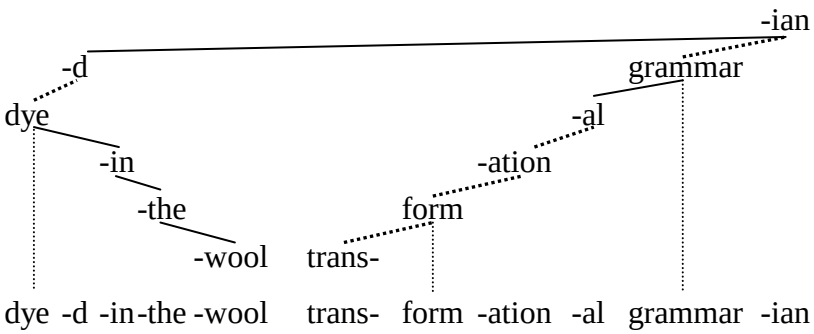
(16)



An analysis along the lines of (16) is parsimonious and simple. No semantic rules are necessary because the pertinent meanings can be read directly off the tree structure. For instance, (16a) shows the relevant catenae *histor-ic-al*, *novel*, *histor-ic-al novel*, and *novel-ist*; these catenae all combine in a straightforward fashion to yield *histor-ic-al novel-ist*. The same is true for (16b), and, for that matter, for all personal noun expressions outright.

Furthermore, the distinction between deep and surface structural representations also loses much of its motivation, since everything that needs to be shown is present on the surface. And the problem with attributive modifiers on the person suffix disappears:

(17)



The complex attributive *dyed-in-the-wool* modifies the person, rather than the grammar. The attributive expression exhibits a further application of the catena, namely one in which the free morphs *-in*, *-the*, and *-wool* recursively cliticize to the morph *dye*. On the catena approach, the fact that the participle morph *-d* intervenes in the linear dimension is irrelevant. Since the morphs *dye*, *-in*, *-the*, and *-wool* form a continuous morph catena, their semantic coherence is guaranteed. The fact that the subordinated morphs lose their ability to constitute prosodic words on their own is symbolized by the hyphens. A hyphen on a free morph indicates that this morph behaves similar to an affix in a specific context.

This section shows how complex words can be analyzed as morph catenae. Once one acknowledges catenae in word structure, a parsimonious account of bracketing paradoxes becomes possible. A catena-based account can stay on the surface, and even allow for a surface-based description of the semantic relationships that motivate the structure.

Bibliography

Anderson, S.R. 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Bauer, L. 1990. Be-heading the word. *Journal of Linguistics* 26. 1-31.

Beard, R. 1991. Decompositional Composition: The Semantics of Scope Ambiguities and

"Bracketing Paradoxes". *Natural Language and Linguistic Theory* 9. 195-229.

Becker, T. 1993. Back-formation, cross-formation, and 'bracketing paradoxes' in Paradigmatic Morphology. In Booij, G. & van Marle, J. (eds.), *Yearbook of Morphology* (vol. 6). Dordrecht: Foris Publications. 1–25.

Di Sciullo, A.M.. 2005. *Asymmetry in Morphology*. Linguistic Inquiry Monographs 46. Cambridge: MIT Press.

Di Sciullo, A.M and E. Williams. 1987. *On the Definition of Word*. MIT: MIT Press.

Embick, David. 2003. Linearization and local dislocation: Derivational mechanics and interactions. *Linguistic Analysis* 33/3-4. 303-336.

Embick, David and Philip Noyer. 2001. Movement operations after syntax. *Linguistic Inquiry* Vol 32, No. 4, 555–595

Embick, David and Rolf Noyer. 2007. Distributed Morphology and the Syntax/ Morphology Interface. Ramchand, Gillian & Charles Reiss eds. *The Oxford Handbook of Linguistic Interfaces*. 289-324. Oxford University Press.

Groß, T. 2010. Chains in syntax and morphology. In *Proceedings of the 24th Pacific Asia Conference on Language*, Information and Computation at Tohoku University, eds. O. Ryo, K. Ishikawa, H. Uemoto, K. Yoshimoto & Y. Harada, 143-152. Tokyo: Waseda University.

Groß, T. and T. Osborne. 2009. Toward a practical DG theory of discontinuities. *Sky Journal of Linguistics* 22. 43-90.

Groß, T. and T. Osborne. To appear. Constructions are Catenae. *Cognitive Linguistics*.

Halle, Morris and Alec Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In Kenneth Hale and S. Jay Keyser eds., *The View from Building 20*, 111-176, Cambridge: MIT Press.

Harley, Heidi and Ralph Noyer. 2003. Distributed Morphology. In *The Second GLOT International State-of-the-Article Book*. Berlin: Mouton de Gruyter. 463-496.

Hudson, R. 1987. Zwicky on Heads. *Journal of Linguistics* 23, 109-132.

Kahane, S. 2003. The Meaning-Text Theory. V. Ágel et al. (eds.), *Dependency and valency: An international handbook of contemporary research*, vol. 1, 546-569. Berlin: Walter de Gruyter.Matthews, P.H. 1972. *Inflectional Morphology: A Theoretical Study Based on the Aspects of Latin Verb Conjugation*. Cambridge: Cambridge University Press.

Mel'čuk, I. 1988. *Dependency syntax: Theory and practice*. Albany: State University of New York Press.

Mel'čuk, I. 2003. Levels of dependency in linguistic description: concepts and problems. Ágel, V. et.al eds. *Dependency and valency: an international handbook of contemporary research*, vol. 1, 188-229. Berlin: Walter de Gruyter.

Mel'čuk, I. 2009. Functional Linguistic Models: A Step Forward in the Study of Man. *International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages, and Their Application to Emergencies and Safety Critical Domains*. Besançon: Presses Universitaires de Franche-Compté, 1-10.

Müller, S. 2003. Solving the bracketing paradox: an analysis of the morphology of German particle verbs. *Journal of Linguistics* 39. 275–325.

O'Grady, W. 1998. The syntax of idioms. *Natural Language and Linguistic Theory* 16:79-312.

Osborne, T. 2005. Beyond the constituent: A DG analysis of chains. *Folia Linguistica* 39(3-4). 251-297.

Osborne, T. in press. Edge features, catenae, and dependency-based Minimalism. *Linguistic Analysis*.

Osborne, T., Michael Putnam, and Thomas Groß. in press.a. Catenae: Introducing a Novel Unit of Syntactic Analysis. *Syntax*.

Osborne, T., Michael Putnam, and Thomas Groß. in press.b. Bare Phrase Structure, Label-less Trees, and Specifier-less Syntax: Is Minimalism Becoming a Dependency Grammar? *The Linguistic Review*.

Pesetsky, D. 1985. Morphology and logical form. *Linguistic Inquiry* 16:193-246.

Pullum, G. 2003. Phrases for lazy writers in kit form. *Language Log*. October 27, 2003. http://itre.cis.upenn.edu/~myl/languagelog/archives/000061.html

Robins, R.H. 1959. In Defense of WP. *Transactions of the Philological Society*. 116-144.

Spencer, A. 1988. "Bracketing paradoxes and the English lexicon." *Language* 64:663-682.

Sproat, R. 1988. Bracketing paradoxes, cliticization, and other topics: The mapping between syntactic and phonological structure. In Everaert et al. (eds), *Morphology and Modularity.* Amsterdam: North-Holland. 339-360.

Stump, G. T. 1991. A paradigm-based theory of morphosemantic mismatches. Language 67/4. 675-725.

Stump, G. T. 2001. *Inflectional Morphology*: *A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.

Williams, E. 2011. *Regimes of Derivation in Syntax and Morphology*. Routledge.

Zwicky, A.M. 1985. Heads. *Journal of Linguistics* 21, 1-29.

A lexicogrammatical perspective in *encoding* dictionaries —with reference to 'pain' examples in English and in Japanese

Masayuki Hirata

Department of Chinese, Translation and Linguistics,
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
mhirata@student.cityu.edu.hk

Abstract

This paper addresses a problem in the current *encoding* bilingual dictionary models, and argues for the necessity of having a lexicogrammatical perspective in such dictionaries, by discussing a pair of equivalents in English and in Japanese. Because existing *encoding* dictionaries base their models on *decoding* dictionaries, they force users to start from a fixed lexical item. This paper demonstrates why such a lexically bound *encoding* process is not preferable, and argues that a more lexicogrammatical approach is necessary in order to enhance the level of naturalness in target languages. First of all, the paper examines the textual grammar of 'pain' both in English and in Japanese. Secondly, it analyses the predicate information according to types of adjectives in Japanese with reference to corpus data. Thirdly, it establishes a lexicogrammatical unit larger than a single lexical item. These arguments will finally lead to a conclusion that an alternative model is necessary for *encoding* bilingual dictionaries.

Keywords

lexicography, lexicogrammar, phraseology, encoding dictionaries

1 Introduction

There are two types of dictionaries, the *decoding* one and the *encoding* one. *Decoding* dictionaries are those that are mainly used in order to understand meanings of a language. *Encoding* dictionaries are designed for the production purpose, in order to express one's ideas in a foreign language etc. Since the advent of computerised corpora, the general quality of dictionaries has been undoubtedly improved, especially the series of monolingual English

learner dictionaries. Yet the improvement in *encoding* dictionaries is scarce, and often there has been a sense of frustration associated with such dictionaries that they are hardly useful in a practical sense. The following English sentence and its Japanese equivalent show a rather complex process of encoding.

1a I have a headache.
1b (*Watashi-wa*) *atama-ga itai* (*desu*).
[(I-*wa*-TOP[1]) head-*ga*-NOM[2] painful (am).]

At a glance, this process may not look too complicated, in terms of the level of vocabulary, or syntactic structures. However, a close comparison reveals that there are few similarities between the two. The subject in English *I* is obligatory, whereas the subject *watashi* in Japanese should be given a null spellout. The verb used in English is *have*, on the other hand, the Japanese translation uses the copula *desu*, which is not overtly shown either. Most importantly, the experience of 'having a pain in one's head' is encoded by the single noun *headache* in English, while Japanese uses the adjective *itai* (painful) together with the noun *atama* (head), which are linked by a nominative case marker *-ga*. A question is naturally posed whether such linguistic differences can be encoded in contemporary lexicography, particularly in bilingual dictionaries for language learners. In what follows, some of the major linguistic frameworks that relate to this matter are described.

Halliday (1998) extensively discusses the grammar of pain and its related expressions in English. Based on the article, Hori (2006) analyses the same domain in Japanese, and there she notes that many of the English expressions classified by Halliday could be translated by the adjective '*itai* (painful)' in Japanese[3]. Hori also observes the inalienability between 'a pain' and its speaker, and claims that "pain is described inherently from the speaker's point of view", and argues that the grammatical complexity of the clause increases as the inalienability between the subject and the pain decreases. However, this view is by no means new or unique. As early as in 1972, Nishio (1972) classifies Japanese adjectives into two categories, attributive and emotional, and he states that emotional adjectives (which include sensory adjectives under its subcategory) take only first person as their subject. Otherwise, emotional adjectives would require a structural change, such as the addition of *-garu* (verbal suffix) or *-souda* (modal marker) in the predicate (Fukuhara 2009, 221-222). Teramura (1983, 1993) also discusses the emotional adjectives quoting Koyama's (1966) observation. According to Teramura, if third person subject is used with emotional adjectives, a variety of modal markers are added to change the structure. He also notes that structural changes do not occur in the past tense and in narratives even when third person subjects are used with emotional adjectives.

Although Halliday and Hori give a comprehensive view of the grammar of pain in English and in Japanese, that is not the aim of the current research. This paper tries to discuss the pair of equivalents more in depth, from the lexicographical point of view. This study focuses on the following points: Firstly, although many linguists claim that emotional and sensory adjectives take only first person subjects, their positions are often unclear whether they assume a null spellout or an overt spellout for such subjects. This has an important implication for the textual grammar of pain. This paper takes the view that sensory adjectives always take a null subject, and filling the slot by any subject (including first person) triggers structural changes. Some corpus evidence will be provided. Secondly, there has hardly been a

[1] TOP: topic marker.
[2] NOM: nominative case marker.
[3] 21 out of 24 examples from Halliday are translated by the adjective *itai*.

consensus on what should be categorised as 'sensory adjectives'. A clearer definition of sensory adjectives will be given in the following sections. Thirdly, based on the discussion, this paper proposes a lexicogrammatically unified unit larger than a single lexical entry. Finally, followed by these arguments, an alternative entry model is proposed for *encoding* bilingual dictionaries.

2 Methodology and data from corpus

The structure of the Japanese translation quoted in 1b will be extensively discussed by using data from a corpus. The corpus used in this research is 'Longman Contemporary Japanese Corpus (LJC)'. It was constructed in 2005 for the creation of *Longman English-Japanese Dictionary* (2007). It has 50 million words, and its written component has 40 million words, divided into the following four categories: academic books/papers, newspapers, magazines, and fictions (each category consists of 10 million words). The spoken component has 10 million words, divided into two categories: A corpus of spontaneous Japanese developed by National Institute of Japanese Language (7 million words), and the captioned data for TV talk shows provided by NHK (Japan Broadcasting Corporation).

2.1 Textual grammar of pain

Halliday (1998) argues that having *I* as a Theme is more natural than having *my head* as a Theme.

> In English (as in many other languages, though not all), there is a particular meaning associated with first position in the clause. Whatever element is put in initial position is being construed by the speaker as the theme of the message: … Now if I say *my head aches*, or *my head's aching*, the first element in that clause is *my head*; I have constructed a message in which my head is presented as the Theme. But this is not the way the situation presents itself to me. Where I start from, what I feel to be the setting of this unpleasant experience, is not my head, it is me—my self, as a whole.
>
> (Halliday 1998, 4)

Halliday further argues that typologically speaking, this is true in many languages. He quotes examples from French, Russian, and Chinese. For example, in Chinese, the language prefers *wǒ tóu téng.* (me+head+aches) to *wǒdi tóu téng.* (my head aches).

Halliday's view should be also reflected in Example 1b. It could be hypothesised that the translation favours *atama-ga itai* (head-*ga* painful) over a noun, for example *zutsuu* (headache), in order to keep the Thematic position to *I*. However, interestingly, it is more natural to have a null subject here than an overt one. There are 108 examples altogether for the pattern *atama-ga itai* (head-*ga* painful) in LJC, nonetheless, there are only 5 examples with the overt subjects, and only 2 of them take first person as a subject, either '*watashi*' or '*ore*'[4]. When the Theme position is actually filled with the subjects, the grammatical structures are somehow affected. Below are the examples quoted from the corpus.

[4] '*Ore*' is first person male singular subject.

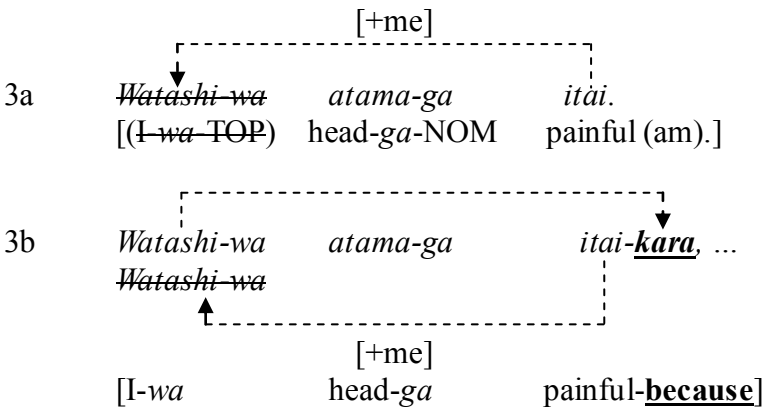
2a *Ore-wa atama-ga itai-**kara**, …*
[I-*wa* head-*ga* painful-**because**]
2b *Watashi-wa atama-ga itaku **nattekita**.*
[I-*wa* head-*ga* painful-**becoming**.]
2c *…Okano wa atama-ga itaku **naru**.*
[*…Okano-wa* head-*ga* painful-**become**.]
2d *Kare-wa shikirini atama-ga itai **to** uttaeteita.*
[He-*wa* constantly head-*ga* painful **that** claimed]
2e *"imouto-wa 'atama-ga itai' **to** sagyou wo yasumitagatta."*
["My sister-*wa* 'head-*ga* painful' **that** she wanted to stop the work."]

Example 2a is marked by the particle *-kara* which is used to give a reason, and the whole sentence was made into a subordinate clause. In 2b, *itai* is changed into a verb form with *-te kuru*, showing a gradual change of the state. In 2c, *itai* is also verbalised with *-naru* to mean 'become painful' without any gradual sense. The quotation particle *-to* is added in 2d and 2e, making the sentence into indirect or direct quotation.

2.2 A link to a null theme

The examples in the previous section empirically supports Teramura's (1983, 1993) arguments. Teramura expands Koyama's (1966) claim and demonstrates several important observations regarding emotional adjectives. First of all, he writes that emotional adjectives with third person subjects can take a variety of modal markers such as *-youda, -souda, -rashii, -hazuda* etc. Secondly, the structural change could happen in subordinate clauses such as giving reasons as in 2a, or reporting clauses as in 2d and 2e, besides in noun modifying clauses which was pointed out by Koyama. Thirdly, he also argues that third person subjects can take the adjective without any structural change in the past tense, because in the past tense, a speaker (or a writer) is not expressing the subject's feeling but simply describing an event in the past. Finally, he adds that in narratives, third person subjects can occur freely with emotional adjectives because of the writing style.

It is worth noting that Teramura does mention that first person subjects tend to be given a null spellout with emotional adjectives. However, he does not anticipate any structural change with an overt first person subjects as in 2a and 2b above. What is interesting here is the fact that the subject being first or third person does not matter much, but any subject which fill the Theme position causes a structural change. A hypothetical structure for the target sentence is described as follows.



The predicate *itai* has the [+me] feature as described in 3a, which has the link to the null subject ~~*watashi-wa*~~. Because of the link, the default setting for the Theme is to be given a null spellout. When the slot is actually filled by other subjects (or even by *watashi-wa* itself), in other words, when the Theme is given an overt spellout, it will cause a further structural change in order to shift the focus of the clause. In 3d, when *watashi-wa* is given an overt spellout, another link has to be shown, in the context of 'giving a reason' in order to differentiate from the old link which connects the adjective and the null theme. The important point here is that the null Theme ~~*watashi-wa*~~ does exist, but it is given a null spellout. Hori (2006) gives a different interpretation as follows:

4	*Atama-ga*	*itai*.
	[Head-*ga*-NOM	painful.]
	(Carrier	Process: relational: attributive)

Here she analyses the body part as a carrier, the grammatical subject of the clause, therefore the Theme would fall onto the body part rather than *I*. This contradicts with Halliday's view. Hori also quotes an example that has *watashi* as an overt Theme. She analyses it as a case of contrastive -*wa* and treat it as a special case (Hori 2006, 215).

> The explicit appearance of the speaker with the particle *wa*, [as in the example] usually marks the start of new information about the speaker or a presentation of him/herself as a contrast to someone else.

However 2a and 2b quoted above (or even 2c to 2e) are clearly not the cases of contrastive -*wa*. Hori's position is unclear as to accepting the existence of null Theme or not.[5] The analysis here takes the view that the null Theme is always present. Also, it assumes the existence of lexicogrammatical link between the predicate and the Theme as depicted in 3a and 3b. The evidence for the null Theme and the lexicogrammatical link will be discussed and presented further in the following sections.[6]

3 Types of adjectives in the predicate

A link to a null Theme, with [+me] feature, is hypothesised in the previous section. This section considers the predicate information more in detail according to the type of adjectives.

3.1 Definition of sensory adjectives

Adjectives in Japanese are semantically classified into three categories, attributive[7], emotional, and sensory. Often the latter two are grouped together as 'emotional/sensory' adjectives. Koyama (1966), Nishio (1972), Masuoka and Takubo (1992) all classify adjectives

[5] Example 4 clearly contradicts with the analysis that anticipates the null Theme. Also, in Hori's (2006, 224) conclusion, she treats 'X-*ga itai*' and '*Watashi-wa* X-*ga itai*" as different structures, therefore, she does not seem to take the view with the null Theme.

[6] Hori presents the view of inalienability between the adjective *itai* and its sensor, however, she does not explain why they are inalienable.

[7] This is the translation of '属性形容詞' (*zokusei keiyoushi*). This is irrelevant to the grammatical distinction such as 'attributive use' or 'predicative use' of adjectives.

into two, including sensory adjectives under the emotional adjectives. Below are the examples of the three types of adjectives with the English translations.

	Example
Attributive adjectives	*ooki* (big), *chiisai* (small), *nagai* (long), *mijikai* (short), *takai* (high), *hikui* (low), *atarshii* (new), *furui* (old), *jyouzu da* (good at), *heta da* (not good at), *chikai* (close), *tooi* (far)
Emotional adjectives	*ureshii* (happy), *kanashii* (sad), *hoshii* (want), *sukida* (like), *nikui* (hate), *sabishii* (lonely), *tanoshii* (fun), *kurushii* (hard), *kawaii* (cute), *hazukashii* (embarrassed), *shinpaina* (worried)
Sensory adjectives	*itai* (painful), *kayui* (itchy), *darui* (dull, tired), *kusuguttai* (tickling), *kusai* (smelly), *urusai* (noisy), *mabushii* (dazzling), *oishii* (tasty)

Table 1 : Semantic classification of Japanese adjectives

Attributive (descriptive) adjectives are used to describe things or people and there is no restriction on which subject can appear with these adjectives. On the other hand, as observed in the previous section, emotional and sensory adjectives could only take first person as their subjects. These adjectives describe one's inner state, therefore, they are subjective.

There seems to be hardly any consensus on what should be included under the category of sensory adjectives. Under Table 1, adjectives which are connected with physical senses of touch, smell, taste, hearing, and seeing are included. In the next section, these sensory adjectives are analysed by the complements they take, the preceding nouns that are marked by the nominative case marker *-ga*. Based on the analysis, a revised definition of 'sensory adjectives' will be given in 3.3 below.

3.2 Types of complements marked by *-ga* for sensory adjectives

Table 2 summarises the types of complements which the sensory adjectives take with the nominative case marker *-ga*. They are ordered according to the frequency. It is immediately visible from the table, that the adjectives *itai* (painful), *darui* (tired, dull), and *kayui* (itchy) almost exclusively take nouns which denotes a body part in their complement position. In the case of *itai* (painful), 486 instances out of 527 occurrences (92.2%) take a body part in its complement position, and the body parts have 24 distinctive types. Similarly, *darui* (tired, dull) takes 8 kinds of body parts, reaching up to 90.6% in its frequency. *Kayui* (itchy) also takes 13 kinds of body parts, exceeding 83%. These figures clearly demonstrate the strong bond between these sensory adjectives and their complements, i.e. a part of body. *Kusuguttai* (tickling) gives a slightly different view. Its overall occurrence in the corpus is 40 times, however, it doesn't take the complement with *-ga* much, unlike the adjectives quoted above. It seems that the feeling of being tickled needs not specify the location in the body.

The adjectives in the right column in Table 2 display a totally different picture. It is very difficult to categorise the words which occur in their complement positions, therefore, only the top five frequent ones are quoted in the table. *Kusai* (smelly) indeed takes a wide variety of complements. It is obvious from the fact that the most frequent one occurs only 3

times out of 41, which is only a little more than 7%. *Urusai* (noisy) and *mabushii* (dazzling) are slightly easier to categorise, *urusai* having a type of sound or voice in its complement position (25.6%), and *mabushii* having a source of light (25%). *Shinryoku-ga mabushii* is more figurative and idiomatic. It should be noted that *oishii* (tasty) takes the comparative marker *hou* most frequently (14.6%), which reveals its inherent nature of being comparative.

	Words before *-ga*	**Freq.**
itai (painful)	BODY PARTS (24 kinds)	486 (92.2%)
	no (nominalization marker)	11
	hou (comparative marker)	7
	kokoro (mind)	7
	kimochi (mind, feeling)	6
	doko (question word)	4
	kore (demonstrative)	3
	koko (here)	3
	TOTAL	**527**

	Words before *-ga*	**Freq.**
darui (tired, dull)	BODY PARTS (8 kinds)	29 (90.6%)
	shita (below)	1
	zentai (whole)	1
	ane (sister)	1
	TOTAL	**32**

	Words before *-ga*	**Freq.**
kayui (itchy)	BODY PARTS (13 kinds)	20 (83.3%)
	tokoro (place)	1
	soko (there)	1
	kyoku (music)	1
	mawari (surrounding area)	1
	TOTAL	**24**

	Words before *-ga*	**Freq.**
kusuguttai (tickling)	BODY PARTS (1 kind)	1
	kokoro (mind)	1
	no (nominalization marker)	1
	TOTAL	**3**

	Words before *-ga* (top 5)	**Freq.**
kusai (smelly)	*iki* (breath)	3
	hou (comparative marker)	3
	-sui (-water)	2
	sore (demonstrative)	1
	kore (demonstrative)	1
	TOTAL	**41**

	Words before *-ga* (top 5)	**Freq.**
urusai (noisy)	*oto* (sound)	11
	koe (voice)	5
	nakigoe (cry of animals)	4
	watashi (I)	2
	kuchi (mouth)	2
	TOTAL	**78**

	Words before *-ga* (top 5)	**Freq.**
mabushii (dazzling)	*shinryoku* (fresh green)	6
	hikari (light)	5
	taiyou (sun)	3
	hizashi (sunlight)	3
	sugata (figure)	2
	TOTAL	**44**

	Words before *-ga* (top 5)	**Freq.**
oishii (tasty)	*hou* (comparative marker)	54
	kore (demonstrative)	15
	kuuki (air)	11
	no (nominalization marker)	8
	shokuji (meals)	6
	TOTAL	**369**

Table 2: Complements of sensory adjectives

3.3 Revised definition of sensory adjectives

A clear distinction should be drawn between the adjectives in the left column (*itai, darui, kayui, kusuguttai*) and the ones in the right column (*kusai, urusai, mabushii, oishii*). Although structurally, all these adjectives can take a complement marked by *-ga* in their preceding position, the types of complements are totally different. To be more precise, the adjectives on the left take almost exclusively the body part as their complements (except *kusuguttai*), whereas a wide variety of words are allowed for the adjectives in the right

column. From another perspective, the source of stimuli which cause the sensations denoted by these adjectives, is inside (or on) ones' body in the case of *itai, darui, kayui, kusuguttai*, while stimuli clearly come from the outside in the case of *kusai, urusai, mabushii, oishii*. Based on these observations, the 'sensory adjectives' should be defined as the adjectives which the source of stimuli come from one's body. This definition still leaves a room to include the adjective such as *kusuguttai* although its complement does not show any similar feature as the other three adjectives. In fact, only a limited set of adjectives is categorised as sensory adjectives in Japanese.[8]

3.4 A particularly amenable structure for 'a pain'

When the scope is narrowed to the adjective *itai*, the fact that 92.2% of complements with *-ga* take a noun which describes a body part could be the strong evidence for the [+me] feature described in 3a and 3b. Because *itai* almost always take one's own body part as its complement, it is nearly redundant to start the sentence with *watashi-wa*. There could be almost no chance for the pain to belong to somebody else. This structure, '(~~*watashi-wa*~~) X-*ga itai*.' is particularly amenable to encode the experience of pain, because the structure is restrictive and productive simultaneously. It is restrictive in a way that it could only encode one's own sensation. This gives no ambiguity, and also, this structural preference prevents other things (for example, a body part, or a pain itself) from taking over the thematic position. In other words, as long as taking this structure to express 'a pain', the thematic position is reserved for '*I*' (*watashi-wa*). It is also very productive since the complement could take almost all kinds of body parts (24 kinds are evidenced in the corpus). This flexibility is much more useful than the noun forms *-tsuu* (-ache), because the noun that could combine with *-tsuu* is limited, as in the case of English *-ache*. Productivity is much lower.

4 *Atama-ga-itai* as one unit

The previous section concluded that sensory adjectives such as *itai, darui,* and *kayui* have a strong bond with their complements, which are almost exclusively a noun that describes a part of body. Based on the data, this section will take a view that *atama-ga-itai* is, in fact, one unit.

Noda (1991) analyses the structure *X-wa Y-ga Z-da*[9] into the following patterns.

(i) When the structure has other case markers such as *-o* (accusative case), *-ni* (dative case), *-no* (genitive case), one of the case markers could be upgraded to a Theme. For example, *Watashi-no kami-ga nagai*. (My hair is long.) could be transformed into *Watashi-wa kami-ga nagai*. (I have long hair.), upgrading the genitive case *Watashi-no* (My) into a Theme *Watashi-wa* (I).

(ii) When two nominative cases are present, one of them could be upgraded into a Theme. The example would be *'watashi-ga kare-ga sukina (koto)'* in which the second *-ga* is requested by the emotional adjective '*suikida*' (like). As in the first case, the first nominative case marker could be upgraded to *-wa* resulting in *'watashi-wa kare-ga sukida.'*(I like him.)

---

[8] The fact that '*kusai, urusai, mabushii, oishii'* allows topicalization taking the topic marker *-wa* clearly shows the difference between this set with the other, i.e. '*itai, darui, kayui, kusuguttai'*. While it is possible to make a general statement '*Taiyou wa mabushii* (Sunshine is dazzling.)', the adjectives in the latter set cannot take *-wa*. Subjectivity must be involved to disallow this topicalization.

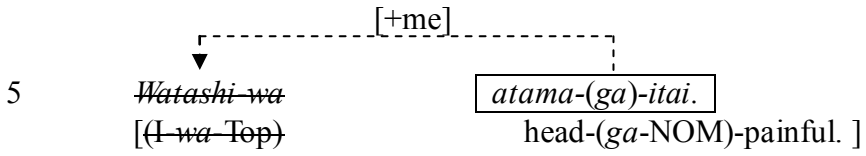
[9] This is the structure which is the focus in this study, assuming the presence of null Theme, and Z takes an adjective.

(iii) There are cases when the adjective and the complement marked by *-ga* forms an idiomatic unit such as '*ashi-ga-hayai*', meaning 'a fast runner'. '*Watashi-ga ashi-ga-hayai (koto)*', could be transformed into '*Watashi-wa ashi-ga hayai*', changing the first *-ga* into a Theme.

'*Atama-ga itai*' should not be explained by (i) above. The underlying structure is not *Watashi-no atama-ga itai*. (My head is painful.). Since the adjective *itai* has [+me] feature, using the genitive case is certainly redundant. When the adjective is an attributive (descriptive) adjective, for example *nagai* (long) in (i) above, it is natural to assume the existence of genitive marker; because the attributive adjective does not have the inherent [+me] feature in itself. In other words, for example in *Watashi-no kami-ga nagai*. (My hair is long.), it is necessary to specify whose hair it belongs to. Also, typologically speaking, as explained by Halliday, the natural Theme of 'having a pain' should be '*I*', rather than 'a body part', therefore, *Watashi-no atama-ga itai*. (My head is painful.) is not a preferable option.

Because of the data presented in the previous section, i.e. the strong bond between *itai* and its complements (a body part), this paper takes the view that *atama-ga-itai* is actually one unit. As described in (iii) above, it is nearly an idiomatic expression. Another empirical evidence that supports this view is that the nominative case marker *-ga* could be dropped in spoken forms, yielding *atama-itai*. *Atama-itai* is a perfectly natural expression in daily speech.

This view also conforms with the idea that the underlying form of *atama-ga itai* is not '*Watashi-no atama-ga itai* (My head is painful).' with the genitive case marker. Because *atama-ga-itai* is actually one unit, it cannot be broken down into two pieces as in [*Watashi-no atama-ga*][*itai*.], as a subject and a predicate. The final output of Example 1b should look as follows:



5 Conclusion

When example 1a is presented, users of *encoding* bilingual dictionaries will be most likely to look for an equivalent of 'headache' in Japanese, therefore, forced to use a noun form, '*zutsuu*', in their translation. This lexically bound *encoding* process is not preferable because such a process completely lacks lexicogrammatical perspective. Often a unit of lexical item could be larger than a single word, and it could have a further link, for example to a Theme, as observed in the case of Japanese adjective, *itai*. A future *encoding* bilingual dictionary should incorporate such information described in Example 5 above, as a part of the adjective headword '*itai*'. In order to do so, a systematic and comprehensive perception of phraseology will be necessary, which makes use of the corpus as a data source both as evidence and statistical interpretation. The current study obviously has the limitation and the application of the theory to a practical lexicography has a long way to go, however, addressing a question on the current *encoding* model for bilingual dictionaries should be regarded as a significant step. *Encoding* dictionaries should not be only a reversed version of *decoding* dictionaries. Not only the lexicogrammatical perspective, but there are also a lot of possibilities to be pursued, such as having a conceptual headword showing a network of usage beyond parts of speech, indicating polarity, probability, and so on.

Bibliography

Fukuhara, S. 2009. Syokyu Nihongo ni dounyusareteiru keiyoushi no bunseki (An analysis of adjectives introduced in elementary Japanese textbooks). In Tomimori, N., M. Minegishi, & Y. Kawaguchi (eds), Koopasu o mochiita gengo kenkyuu no kanousei (A possibility of corpus based language studies), 219-248. Tokyo: Tokyo University of Foreign Studies.

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar,* 2nd edn, London: Edward Arnold.

Halliday, M. A. K. 1998. On the grammar of pain, *Functions of Language,* 5 (1): 1-32.

Halliday, M. A. K. & C. M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar,* 3rd edn, London: Hodder Arnold.

Hori, M. 2006. Pain expressions in Japanese. In Thompson G. & S. Hunston (ed). *System and Corpus,* 206-225. London: Equinox Publishing Limited.

Koyama, A. 1966. -no -ga -wa no tsukaiwake ni tsuite (Different usage of -no -ga -wa). *Kokugogaku* (Japanese Linguistics), 66.

Longman Dictionaries. 2007. *Longman English-Japanese Dictionary.* Tokyo: Pearson Education.

Masuoka, T. & Y. Takubo. 1992. Kiso Nihongo Bunpou (Basic Japanese Grammar), 2nd edn, Tokyo: Kuroshio.

Nishio, T. 1972. Keiyoushi no imi, youhou no kijyutsu teki kenkyu (Descriptive analysis of meaning and usage of adjectives), Tokyo: Shuei Shuppan.

Noda, H. 1991. Wa and Ga (Particle –Wa and –Ga), Tokyo: Kuroshio.

Teramura, H. 1982. Nihongo no Shintakkusu to Imi (The Syntax and Semantics of Japanese), Vol. 1. Tokyo: Kuroshio.

Teramura, H. 1993. Teramura Hideo Ronbunshu (The Collected papers of Hideo Teramura), Vol. 2. Tokyo: Kuroshio.

Valency Ambiguity Interpretation: What Can and What Cannot be Done

Boris Iomdin (1), Leonid Iomdin (2)

(1) Theoretical Semantics Department – Institute of Russian Language, Russian Academy of Sciences
Moscow, Russia
iomdin@ruslang.ru
(2) Laboratory of Computational Linguistics – Institute for Information Transmission Problems, Russian Academy of Sciences
Moscow, Russia
iomdin@iitp.ru

Abstract

Russian constructions that involve the ambiguity of valencies are considered regarding the extent in which it can be successfully resolved by man or machine. The material includes two types of phenomena: 1) Russian counterparts of noun phrases like (a) *the phases of sleep* vs. (b) *the phase of active sleep* – in (a), *sleep* instantiates the subject valency of *phase* whereas in (b) *sleep* is the content of the *phase*; 2) subject and object infinitives with the verbs *prosit'* 'ask' and *predlagat'* 'suggest/offer': *Rebënok prosit est'* lit.'The child asks to eat' vs. *Rebënok prosit podojti* lit.'The child asks (for someone) to come up', *On predložil vstretit menja* 'he offered to meet me' vs. *On predložil prijti k nemu* 'he suggested that (I) should come round to him'.

Keywords

Valency structure, automatic disambiguation, human interpretation of texts, surface syntax, deep syntax, semantics, lexicography

1 Introductory Remarks

Any language has certain constructions that are ambiguous with regard to how actants of situations represented are expressed. A classical case is subject/object ambiguity like *support of the government* ('someone supports the government' vs. 'the government supports someone') or *the betrayal of her husband* (who betrayed whom?). This phenomenon is much less common in English (because composite constructions like *government support* clear up some of the ambiguities) than in Russian, where it is fairly widespread, or in Latin, which

provided the famous example *amor patriae* ‘love of one’s fatherland’ for Alan Gardiner (1932). Gardiner considered this ambiguity in detail, postulating two genitive cases: in one of the interpretations the word-form *patriae* is in *genetivus objectivus* and points to the fatherland as object of love whilst the other interpretation uses *genetivus subjectivus* and refers to the fatherland as the loving entity. Despite the fact that it is not always possible to resolve this ambiguity, especially in automatic text processing tasks, it is quite lucid, both logically and semantically, and presents no theoretical difficulties.

A quite different case of actant ambiguity is seen in Russian constructions like

(1) *na vysote Monblana* ‘at the height of Mont Blanc’,

which may refer to the mountain itself and its height, or to a different place whose height is equal to the height of this particular Mont Blanc (imagine e.g. a plane flying at such height). Juri Apresjan (1974) believes that such examples provide another instance of actant ambiguity where oscillation occurs between the parameter argument (whose height is assessed) and its value (how high it is): he sees a parallelism between the second interpretation of (1) and expressions like *na vysote 4810 metrov* ‘at the height of 4810 meters’). The semantic opposition in this type of ambiguity is quite clear, too (even though its theoretical character appears to be different, which will be discussed later).

This paper is focused on two other cases of actant ambiguity, where the actants to be interpreted are related in a less trivial way. The first case (Section 2) refers to surface syntax as understood in Igor Mel’čuk’s Meaning ⇔ Text Theory; the second (Section 3) refers to DSyntR or even SemR representation.

2 Ambiguities of Self-Derivatives and Words of Close Lexical Classes

2.1 Material

Actant ambiguity regularly emerges in phrases that contain nouns with a valency of content instantiated by the genitive. Most such nouns are self-derivatives (Russian «avtoderivaty», see Boguslavsky & Iomdin, 2010a,b): a typical example is *ideja* ‘idea’, which has two indisputable valencies,[1] the subject (i.e. the originator of the idea) and the content (i.e. what the idea is about): cf. resp. (2) *ideja Ejnštejna* ‘Einstein’s idea’ and (3) *ideja otnositel’nosti vremeni i prostranstva* ‘the idea of space and time relativity’. Obviously, many phrases with *ideja*, including (2) and (3), are unambiguous with regard to its valencies of subject and content as the words filling them are categorially different: prototypical instantiations of the subject are names of humans whilst prototypical actants of content are propositions or facts. The real difficulties start when these two valencies are instantiated in a less prototypical way.

[1] The noun *ideja* may have another semantic valency expressible by genitive, that of theme, as in *osnovnye idei jadernoj fiziki* ‘the basic ideas of nuclear physics’, which should probably be distinguished from both the valency of content and the valency of subject. This was suggested e.g. by Elena Uryson (2004: 254) in her dictionary entry for *ideja*. For simplicity’s sake, we will abstract away from this valency.

For instance, the content of an idea may be expressed by a word denoting an organization, as in

(4) *ideja naučno-issledovatel'skogo universiteta* 'the idea of a research university'

or even a human, as in

(5) *My videli, kak trudno bylo moskovskim umam osvoit'sja s ideej vybornogo carja* (V. Ključevskij) 'We saw how difficult it was for the Moscow minds to accept the idea of an elected tsar'.

Cases like (4) or (5) offer a metonymical, condensed description of a situation: (4) roughly means that somebody had an idea to set up a research university while (5) introduces the idea that the Tsar should be elected.

Conversely, the subject valency may often be filled by a noun naming a collective author: a science, theory, literary work etc, as in *idei jadernoj fiziki* 'the ideas of nuclear physics'. In such cases, which immediately become ambiguous, no categorial distinction of the actants is possible. They will be our main focus of interest.[2]

The phenomenon we are discussing appears to involve broad linguistic material. Russian has hundreds, if not thousands,[3] of nouns that have at least two valencies fillable by the genitive. The examples below are given to demonstrate the vast range of semantic classes of nouns which have this feature and may trigger potential syntactic ambiguities.

(a) ***Tema*** 'theme, topic', ***gipoteza*** 'hypothesis'**,** ***aksioma*** 'axiom'**,** ***lozung/slogan*** 'slogan'**,** ***deviz*** 'motto'**,** ***zamysel*** 'plot' etc. *Tema X-a* 'theme of X' may refer to a theme offered by X or consisting in X, cf. *fantazija na temy Čajkovskogo* 'a fantasia on the themes of Tchaikovsky' vs. *tema vojny* 'the theme of war', *gipoteza Sepira-Uorfa* 'the Sapir–Whorf hypothesis' vs. *gipoteza Boga* 'the hypothesis of God', *lozung partii* 'the party's slogan' vs. *lozung spravedlivosti* 'the slogan of justice'; **(b)** ***čuvstvo*** 'feeling', ***blaženstvo*** 'bliss', ***radost'*** 'joy'**,** ***bojazn'/strax*** 'fear' and many other emotions (though not all of them, e.g. not *revnost'* 'jealousy' and not *gnev* 'wrath, anger'): *čuvstvo materi* 'the feeling of a mother' vs**.** *čuvstvo opasnosti* 'the feeling of danger', *strax <bojazn'> rebënka* 'the fear of a child' vs. *bojazn' vysoty* 'the fear of heights'; **(c)** ***variant*** 'variant', ***versija*** 'version'**,** ***svojstvo*** 'property'**,** ***ottenok*** 'shade (of colour)'**,** ***rol'*** 'role'**,** ***funkcija*** 'function': *versija sledovatelja* 'the version of the

[2] We confine ourselves to considering the cases where ambiguous valencies are instantiated by nouns, thus disregarding interesting constructions in which ambiguous or syncretic realizations of nominal valencies by adjectives take place: *evropejskie cennosti* 'European values' (Europe is the subject), *russkaja ideja* 'The Russian idea' (Russia or the Russians may be the subject and/or content of the idea), *xlebnye zagotovki* (grain procurements (grain is the object), *nemeckij plen* 'German captivity' (captivity by the Germans, who are the subject) vs. *nemeckij voennoplennyj* 'German prisoners of war' (Germans who are in captivity, i.e. they are the object).

[3] The combinatorial dictionary of Russian, which has ca. 100,000 entries and is part of the ETAP-3 linguistic processor (Apresjan *et al.* 2003), features almost 1200 nouns whose first and second syntactic valencies can be instantiated by the genitive. Of course, not all of them are of the class with which we are dealing (in particular, the nouns that produce subject/object ambiguity are there, too), but on the other hand this dictionary cannot be considered complete, either with respect to the list of words included or with respect to government patterns of individual nouns.

investigator' *vs. versija man'jaka-odinočki <ubijstva iz revnosti>* 'the version of a lone maniac <of murder out of jealousy>'; *svojstvo sistemy* 'the property of the system' vs. *svojstvo refleksivnosti* 'the property of reflexivity'; *ottenok cveta* 'the shade of colour' vs. *ottenok sinego* 'the shade of blue', *kommunikativnaja funkcija teksta* 'the communicative function of a text' vs. *tekst vypolnjaet funkciju soobščenija* 'the text implements the function of communication'; **(d)** ***reputacija/renome*** 'reputation'**,** ***slava*** 'fame'**:** *Reputacija eksperta byla bezuprečnoj* 'The expert's reputation was impeccable' vs. *On pol'zuetsja reputaciej eksperta* 'He has the reputation of an expert'; **(e)** ***etap/stadija*** 'stage', ***faza*** 'phase', ***period*** 'period'**,** ***stupen'/šag*** 'step' etc.: *etapy stroitel'stva doma* 'the *stages of house construction*' vs. *etap zakladki fundamenta* 'the stage of foundation laying'; *na raznyx periodax razvitija* 'during different periods of development' *vs. period upadka* 'the period of decline'; *faza Luny* 'the phase of the Moon' vs. *faza rastuščej <ubyvajuščej> Luny* 'the phase of the growing <waning> Moon; **(f)** ***sostojanie*** 'state', ***zvanie*** 'rank', ***status*** 'status'**,** ***professija*** 'profession'**,** ***prizvanie*** 'vocation'**,** ***amplua*** '(an actor's) line of business': *sostojanie bol'nogo* 'the state of the patient' vs. *sostojanie affekta* 'the state of affect', *zvanie Ivanova* 'the rank of Ivanov' vs. *zvanie admirala* 'the rank of an admiral'; **(g)** ***akt*** 'act'**,** ***akcija*** 'action'**,** ***demonstracija*** 'demonstration': *akt otčajavšegosja čeloveka* 'an act of a desperate man' vs. *akt samopožertvovanija* 'an act of self-sacrifice'; *demonstracija nedovol'nyx* 'a demonstration of discontented people' vs. *demonstracija nedovol'stva* 'a demonstration of discontent'.

2.2 Criteria for Disambiguation

We know of three general criteria that can be used to resolve actant ambiguity of the type discussed here: 1) context involving lexical functions, 2) adverbial derivatives and 3) instantiation of the respective valencies by pronouns. None of these criteria is absolute but their consideration will allow us to substantially reduce the level of uncertainty when interpreting potentially ambiguous utterances.

2.2.1 Lexical Functions

Information on types of lexical function (LF) whose values are expressed with the argument nouns that have equally realizable valencies has long been successfully used to resolve ambiguity in automatic text analysis of Russian and English. Papers by linguists belonging to the Moscow semantic school describe this in detail (see e.g, Apresjan *et al*. 2007, 2010). Let us see how this information can help the interpretation of actants of words of the *ideja* class.

Considering the LFs listed in the dictionary entry for *ideja*, ones which can be used for disambiguation purposes are the LFs oriented at the valency structure of the key word, such as Oper and Func. On the one hand, these are the LFs of the $Oper_1$ and $Func_1$ classes, which are oriented at the first valency of the word: cf. $Oper_1$(*ideja*) = *imet'* 'to have an idea', S_0Oper_1(*ideja*) = *naličie* 'existence of an idea', $IncepOper_1$(*ideja*) = *vydvigat'* / *vyskazyvat'* 'suggest / put forward an idea', $S_0IncepOper_1$(*ideja*) = *vydviženie* 'suggestion of an idea', $CausOper_1$(*ideja*) = *natalkivat' na* 'lead to the idea', $Func_1$(*ideja*) = *prinadležat' / isxodit' ot / imet'sja u* 'the idea belongs to / comes from smb.', $IncepFunc_1$(ideja) = *pojavljat'sja u / voznikat' u / roždat'sja u / prixodit' k / osenjat' / ozarjat'* 'the idea arises from someone'; on the other hand, these are the LFs of the $Oper_2$ and $Func_2$ classes, oriented at the second

valency; cf. $Oper_2$(*ideja*) = *javljat'sja* ‘the idea is’, Func(*ideja*) = *kasat’sja / zatragivat’* ‘the idea concerns smth.’.

The LF criterion functions in “crisscross” manner. If, in a sentence with *ideja X-a* ‘idea of X’, the word *ideja* is bound to another word within a lexical function of the first class, the subject valency is instantiated with the subject governed by the LF verb, so X must be interpreted as the valency of the **content of the idea**, e.g. its second valency: *Vo sne Dante* ***ozarila ideja*** *«Božestvennoj comedii»* [X] ‘In sleep, the **idea** of “La Divina Commedia” **dawned upon** Dante’. If, however, the word *ideja* is bound to a word within a lexical function of the second class, the content valency is instantiated with an NP governed by the LF verb, so X must be interpreted as the valency of the **subject of the idea**, e.g. its first valency: *Važnoj* ***ideej*** *fil’ma* [X] ***javljaetsja*** *predatel’stvo* ‘An important idea of the film is treachery’.

2.2.2 Adverbials

Many nouns of the class in question can be part of prepositional collocations modifying a verb which instantiates one of the semantic valencies of these nouns: *po idee* ‘according to the idea’, *po pričine* ‘because of’, *s cel'ju / v celjax* ‘with the purpose of’, *na širote* ‘at the latitude of’, *s vysoty* ‘from the height of’, etc[4]. The valency structure of adverbial derivates is quite complicated, and we are not discussing it here[5]. For our purposes, it suffices to say that the noun within such prepositional collocations governs one of its valencies. This fact can be used as a disambiguation criterion: in many of such adverbials, only one of the two types of valencies can be instantiated with a NP in the genitive case governed by the noun.

Let us consider the adverbial *po idee* ‘according to the idea’. The word *ideja* here can only govern its subject valency, as in *Po idee avtora figura pojavljajuščejsja iz morja prekrasnoj damy simvoliziruet roždenie goroda* ‘According to the author’s idea, the figure of a lady appearing from under the sea symbolizes the birth of the city’, where the word form *avtora* instantiates the subject valency of the word *ideja*, and its content valency is instantiated by the rest of the sentence. Outside the adverbial the word *ideja*, as we have seen, can govern both its actants in the genitive case. Similarly, in *lingvisty priexali v Barselonu s cel’ju učastija v konferencii* ‘The linguists have come to Barcelona with the objective of participating in the conference’ the word *cel’* governs the word *učastija* (its second actant), while without the adverbial the word *cel’* can also govern its first actant, as in *Cel’ lingvistov – učastie v konferencii* ‘The linguists’ objective is the participation in the conference’.

For some nouns, it may even be the case that within the adverbial only one of its two valencies can be instantiated, while without the adverbial it can only govern the other valency.

Consider the situation described by the sentence *Podžog privel k požaru* ‘Arson brought about the fire’. Discussing it, we can use the expressions like *pričina požara* ‘cause of the fire’ (but not $^{\#}$*pričina podžoga* ‘cause of the arson’) or *po pričine podžoga* ‘due to the arson’ (but not

[4] The degree of idiomaticity of these adverbials is different. In some of the cases, the adverbial is far in its meaning from the noun which belongs to it and should be considered an independent lexical unit, cf. *sila* ‘force, power’ – *v silu* ‘by virtue of’, *storona* ‘side’ – *so storony* ‘*on the part of*; *protivoves* ‘counterbalance’ – *v protivoves* ‘as the counter to; versus’: we do not consider such adverbials now.

[5] This structure is considered in detail by Igor Boguslavsky (2008).

#po pričine požara ‘due to the fire’); see Boguslavsky & Iomdin 2010a. The same goes for *arest prestupnika* ‘arrest of a criminal’ vs. *pod arestom policii* ‘under police arrest’.

Several more examples of diagnostically significant adverbials and nouns forming them are given below with no detailed comments: ***na puti*** *stroitel'stva* ‘in the way of construction’ – ***put'*** *prob i ošibok* ‘the way of trial and error’, ***pod bremenem*** *obstojatel'stv* ‘under the burden of circumstances’ vs. ***bremja*** *belogo čeloveka* ‘the white man’s burden’, ***po zakazu*** *klienta* ‘by request of the customer’ vs. ***zakaz*** *novyx komp'juterov* ‘order of new computers’*;* ***po zaprosu*** *prokurora* ‘at the request of the prosecutor’ – ***zapros*** *dopolnitel'nyx svedenij* ‘request of additional data’, ***v zaščitu*** *obvinjaemyx* (‘in defence of the accused’ = object) – ***pod zaščitoj*** *gosudarstva* (‘under the protection of the state’ = subject) – ***zaščita*** *advokata* ‘defence offered by a lawyer’, ***v popytke*** *samozaščity* ‘in an attempt to protect oneself’ – ***popytka*** *čeloveka zaščitit’ sebja* ‘the man’s attempt to protect himself’, ***vo slavu*** *pobeditelja* ‘to the glory of the conqueror’– *mirovaja slava etogo skripača* ‘the world-wide fame of this violininst’, etc.

There are two complicating factors. First, some of the adverbial derivates can yet accept different valencies. Consider the noun *pravilo* ‘rule’ that has as many as three valencies instantiated by the genitive: a) subject (≈ the rule’s author), cf. *pravilo Mirandy* ‘the rule of Miranda’; b) object (to which the rule is applied), cf. *pravila dorožnogo dviženija* ‘traffic rules’; and c) content (what the rule consists of), cf. *pravilo levoj ruki* ‘the left hand rule’, *pravilo buravčika* ‘the corkscrew rule’ etc.; cf. also the English *rule of thumb.*[6] This noun has an adverbial derivative *po pravilu* ‘by the rule’, which differs from the adverbials listed above in that it accepts any of the three valencies in genitive: *po pravilu Mirandy* ‘by the rule of Miranda’, *po pravilam dorožnogo dviženija* ‘according to traffic rules’, *po pravilu buravčika* ‘by the corkscrew rule’.

Second, it is not always possible, especially in NLP tasks, to differentiate between an adverbial derivative and a free prepositional phrase, in which the restriction on actant attachment is lifted. Cf.: *Priexali s cel’ju učastija* ‘came with the aim of participating’, where *s celju* is an adverbial, and *Ja soglasen s celju učastnikov* ‘I agree with the objective of the participants’, where *s celju* is a free prepositional phrase.

2.2.3 Pronouns

If valencies of our nouns are filled with pronouns, this can be diagnostically significant for valency disambiguation. We will look at the noun *ideja* again to see what pronouns can fill its subject and content valency slots. It appears that the subject valency is readily instantiated by third person personal pronouns in the genitive *ego, eë, ix* (semantically equivalent to *his/her/its/their*) or possessive adjectives like *moj* ‘my’, *vaš* ‘your’, *čej* ‘whose’, *čej-to* ‘somebody’s’ etc. We will refer to them as Group A pronouns. As for the content slot, it is not normally filled by Group A pronouns (even though this ban is not absolute, especially for third person pronouns). Instead, it is filled by pronominal nouns in the genitive like *ètogo/sego* ‘of this/thereof’, *čego* ‘of what’, *čego-to* ‘of something’ etc. or pronominal

[6] Interestingly, this valency of “brief content” often presupposes a metaphor impossible to understand without background knowledge.

adjectives *ètot* ‘this’, *kakoj* ‘what, of what kind’, *takoj* ‘such’ and a few others: these will be referred to as Group B pronouns.

Cf.: *ideja Petra* ‘Peter’s idea’, *ego ideja* ‘his idea’, *moja ideja* ‘my idea’, *č'ja-to ideja* ‘somebody’s idea’ vs. *ideja večnogo dvigatelja ne nova* ‘the idea of perpetual motion is not new’, *ideja ètogo ne nova* ‘the idea of this is not new’, *èta ideja ne nova* ‘this idea is not new’, *Kakaja ideja ne nova?* ‘What idea is not new?’, etc. Remarkably, the relative pronoun *kotoryj* ‘which’ is ambivalent in this respect: cf. *učënyj, ideja kotorogo…* ‘the scientist whose idea…’ and *večnyj dvigatel’, ideja kotorogo…* ‘the perpetual motion the idea of which…’ are equally acceptable.

Pronominal instantiation of valencies considered here are of a systemic character. We studied this phenomenon on a large class of nouns and found that in the overwhelming majority of cases our observation holds true. To give an example, out of 45 cases of contact placement of the 3rd person pronoun with the word *avtorstvo* ‘authorship’ found in the National Corpus of Russian, 33 pronouns could be definitely identified as the subject fillers for *avtorstvo*, while the remaining 12 could either be identified as objects or were ambivalent. As was noted, the ban on the second valency is less strict in this case.

On the other hand, pronominal realizations of valencies often display individual peculiarities and largely depend on the semantic class of the noun. For the nouns belonging to the semantic field of images, pronominal filling of slots loses its discriminative power: *moj portret* ‘my portrait’ may readily mean both the portrait made by me and the portrait depicting me.

For words whose slots obey the above regularites, the pronominal criterion could be conveniently used in human disambiguation of valencies: in order to understand which valency is expressed by a noun in an utterance, one may try to re-phrase it with a pronoun and see which of them fits.

To illustrated this claim we will look again at constructions like (1) *na vysote Monblana*. We remember that the parametric word *vysota* ‘height’ has two valencies: the parameter’s argument (carrier) and its value. The first of these valencies is pronominalized naturally by Group A pronouns: *vysota gory* ‘the height of the mountain’ – *eë vysota* ‘its height’, *č'ja vysota* ‘whose height’ etc. Conversely, the second valency is expressible by Group B pronouns: *vysota 8 metrov* ‘the height of 8 meters’; *èta vysota* ‘this height’, *kakaja vysota* ‘which height’, etc., but not *#eë vysota* or *#č'ja vysota.* If we agree with Juri Apresjan’s opinion that in the adverbials like (1) the word *Monblana* fills the value slot of the parametric noun, we will have difficulty explaining why these adverbials can be pronominalized with Group A pronouns: *na ego vysote* ‘at its height’*, s č’ej vysoty* ‘from whose height’, etc. These constructions are quite common in language use. (6) presents a characteristic literary example:

(6) *Džomolungma govorit: “Ja vyše, ibo s moej vysoty ves’ mir viden!” A Golgofa v otvet: “Net, ja vyše, ibo s moej vysoty čelovek kak na ladoni”* (A.Korjakovtsev) ‘Jomolungma says: I am higher since one can see the whole world from my height. But Golgotha replies: No, I am higher, since from my height man is seen as plain as the nose on one’s face’.

These facts convince us that in the expressions like (1) the first valency of the parameter is saturated. As for the fact that expressions like *s vysoty X-a* ‘from the height of X’ or *na vysote X-a* ‘at the height of X’ X may characterize the height of a different object Y (which is

positioned at the height equal to the height of X), it should be interpreted along the same lines as expressions like *ulybka rebënka* ‘the smile of a child’, where the genitive conveys the sense of similarity.

3 Ambiguities of Subject/Object Infinitive

3.1 Material

We will now briefly discuss another remarkable ambiguity case, this time of the deep, rather than the surface, syntactic nature: the “verb – infinitive complement” construction. Most verbs which allow this construction fall into two classes. For verbs of the first class, the subject of the infinitive is coreferent with the subject of the head verb, as in *Ja xoču spat'* lit. ‘I want to sleep’ or *Ona ljubit tancevat'* lit. ‘She likes to dance’. For verbs of the second class, the infinitive refers to the head verb’s object, as in *General velit vystupat'* ‘The general orders [the soldiers] to attack’ or *Učitel' razrešaet otvetit'* ‘The teacher permits [the students] to answer’. For some head verbs, both interpretations are possible.

3.1.1 *Predlagat'* ‘to offer’ vs. ‘to suggest’

First, consider the verb *predlagat'* ‘to offer, to suggest’. It can govern subject infinitives, cf. *Xodil na vokzal, predlagal passažiram pomoč' snesti veščи* (A. Panteleev) [lit. ‘(He) used to come to the railway station and offer the passengers to carry their luggage’]; *Eščë bol’še on obradovalsja, kogda Ženja predložila vymyt' posudu* (O. Novikova) [lit. ‘He was even more glad when Zhenya offered to wash the dishes’]. But it can govern object infinitives as well, cf. *Predlagaem čitateljam razrabotat', izgotovit' i ispytat' takoe prisposoblenie* (B. Sinelnikov) [‘We suggest that the readers work out, manufacture and test such a device’]; *V radiostudii mne predlagajut, poka ja ždu, poslušat' zapisyvaemuju peredaču* (M. Gamburd) [‘In the radio studio they suggest to me that, while waiting, I listen to the show being recorded’].

The most important thing here is that in many cases it is quite hard (or impossible) to determine whether the infinitive refers to the subject or to the object. Cf. *Ol'ga predlagaet otvezti tëte Mane samovar* (G. Shcherbakova) [‘Olga offers to take the samovar to aunt Manya’ OR ‘Olga suggests that someone takes the samovar to aunt Manya’]; *On predložil Mižuevu dat' deneg na eto delo, i Mižuev radostno soglasilsja* (M. Artsybashev) [‘He offered to Mizhuev to give money for the cause, and Mizhuev gladly agreed’ OR ‘He suggested to Mizhuev that Mizhuev give money for the cause, and Mizhuev gladly agreed’].

These two situations may well represent two different senses of the verb *predlagat'*, which could roughly be translated as ‘to offer (to do something)’ and ‘to suggest (that someone else does something)’ (this is the way this verb is described in some Russian dictionaries, e.g. in Malyj Akademicheskij Slovar). Note that the second sense includes semantic components conveying the authority of the subject over the object of the situation and even those of mild compulsion exerted upon the object. Such an interpretation is also supported by the fact that the subject and the object infinitives cannot be coordinated. Indeed, sentence (7), though originating from a novel by Vasily Grossman, a very good author, seems highly infelicitous:

(7) *On daže predložil mne davat' Jure uroki francuzskogo jazyka i platit' za urok tarelkoj supa* 'He even suggested to me to give French lessons to Yura and to pay with a plate of soup for each lesson',

where the author clearly meant that 'me' is to give lessons and 'he' is to pay. But even with this lexical interpretation, one is often faced with the challenge to disambiguate between the two senses of *predlagat'*.

The case where the infinitive governed by *predlagat'* refers to both the subject and the object of the verb simultaneously (which occurs very frequently, cf. e.g. (8) and (9)) deserves special consideration:

(8) *Predlagaju pojti v kakoj-nibud' restoran* 'I suggest that we go to a restaurant';

(9) *Kto-to predložil posmotret' novyj fil'm* 'Someone offered that [everyone] watched a new movie'

On the one hand, such sentences are fully correct and do not constitute any pun, which is generally the case if a word is used in two senses simultaneously (cf. Apresjan 1974: 180-187), so only one sense of *predlagat'* is present. On the other hand, we believe that such uses of *predlagat'* do not represent any additional lexical meaning of the verb and should be considered on a par with the sentences where *predlagat'* accepts a **subject** infinitive. Indeed, sentences like (8) or (9) do not presuppose any authority or coercion of the subject; besides, in many cases, sentences with *predlagat'* and a non-object infinitive defy unequivocal intepretation: in (10), it is not clear whether the subject offers to sing alone or in chorus:

(10) *On predložil spet' novyju pesnju* 'He offered to sing a new a song'.

Such behavior of *predlagat'* is similar to the behavior of the pronoun *we*: despite the fact that it can be used inclusively or exclusively of the hearer, its lexical meaning remains the same.

3.1.2 *Prosit'* 'to ask'

Second, consider the verb *prosit'* 'ask'. Quite unlike *predlagat'* – and, remarkably, in sharp contrast to the English verb *to ask* – in most cases it governs an object infinitive, cf. *Babuška prosit prinesti očki* 'The grandmother asks to bring her spectacles', *Provožajuščix prosjat vyjti iz vagonov* 'People who see passengers off are requested to leave the coaches'. But in a number of cases it can also govern a subject infinitive; cf. *Rebënok prosit pit'* 'The child asks for something to drink'. Here, too, both interpretations of the infinitive may be possible in many cases. E.g. phrases like *prosit' zavtrakat' <obedat', užinat'>* lit. 'to ask to have a breakfast <lunch, dinner>' are used in both senses, cf. *I barina prosit' obedat'!* (A. S. Pushkin) 'Invite the master to lunch!' vs. *Ne žnët i ne kosit, a obedat' prosit* lit. 'He does not reap or mow but he asks for lunch' (a Russian riddle about the idler).

Interestingly, the second construction is only possible with a very limited number of verbs denoting the consumption of food (in a broader sense, including eating, drinking, or smoking): *prosit' est' <kušat', žrat'>* 'to ask for food', *prosit' pit' <popit'>* 'to ask for water', *prosit'*

vypit' ‘to ask for alcohol’[7], *prosit' kurit' <zakurit'>* ‘to ask for cigarettes’. Even synonyms of the above verbs (as the matter of fact, their Aktionsarten) may be unacceptable, cf. ??*Mal'čik prosit napit'sja* ‘The boy is asking for enough water to quench his thirst’, ??*Mal'čik prosit zapit' lekarstvo* ‘The boy is asking for water to take with his medicine’, ??*Mal'čik prosit otpit'* ‘The boy is asking to take a sip’ etc. Verbs denoting other physiological needs are hardly possible here, cf. ??*Rebënok prosit písat'* lit. ‘The child asks for a piss’, **Devočka ustala i prosit spat'* ‘The girl is tired and is asking for sleep’.[8]

Outside this semantic field, the construction is very rare. Compare the following examples with very close synonyms *poprobovat'* and *isprobovat'* ‘to taste, to try out’: (11) is about food and thus almost normal, whereas (12) is not and appears infelicitous:

(11) *Jabloček mal'čiku mama prinosit. Každyj poprobovat' jabločko prosit* [lit. ‘Mom brings some apples to her boy. Everyone asks [to give him] to taste an apple’]

(12) *Potom on poprosil menja isprobovat' mylo na dele i pošël na rečku myt' golovu* (F. Iskander) ‘Then he asked me to try out the soap and went to the river to wash his hair’.

In the construction with the subject infinitive, the head verb *prosit'*, too, may hardly even be replaced by its close synonyms, cf. ??*On kljančil <vymalival> pit'* lit. ‘He begged to drink’.

On the other hand, a few other Russian verbs reveal a similar behavior, in particular, the verb *potrebovat'* ‘require’, *sprosit'* (in the obsolete sense ‘demand or order, like in a restaurant’): *On sprosil užinat' i stal rasskazyvat' ej podrobnosti begov* (L. N. Tolstoy) ‘He asked to have supper and started to relate to her the details of the race’.

In this context, the verb *davat'* ‘give’ deserves special attention. This verb has two major and contrasting senses; one denoting a simple physical transfer of objects (*davat'* 1, cf. *on dal mne xleba* ‘he gave me some bread’) and the other conveying a permission (*davat'* 2, as in *on dal mne otdoxnut'*) ‘he allowed me to take a rest’. Normally, only *davat'* 2 may govern an (object) infinitive whilst *davat' 1* admits neither the object nor the subject infinitive. However, in the context of consumption of food, *davat' 1* accepts the infinitive, too, as in *On dal mne poest'* ‘He gave me to eat’, invoking a case of mimicry: syntactically, it looks exactly like *davat' 2*, which it is not[9]; see a more detailed description of *davat'* in (Iomdin B., in progress).

These facts make us hypothesize that the infinitives of consumption in this context may in fact be equalled to nouns close in meaning to ‘food’, ‘drink’, ‘tobacco’, or ‘alcohol’[10] (cf. Footnote 7), thus creating self-derivatives out of the respective verbs in the strong sense of

---

[7] This expression is especially interesting, because beyond the construction considered the verb *vypit'* (an Aktionsart for *pit'* ‘drink’) does not necessarily imply an alcoholic beverage.

[8] Maybe this restriction explains why the infinitive is not listed as a possible means of instantiating the content valency in the very detailed description of two meanings of *prosit'* by Marina Glovinskaya, who explicitly writes that the lexeme of *prosit'* with the meaning ‘X asks that Y should give P to X’ does not govern an infinitive (Glovinskaya 2004: 885).

[9] This claim is easy to prove: you can well say something like *Levoj rukoj on dal mne popit'* ‘with his left hand he gave me to drink’, which unequivocally points to the physical nature of the action.

[10] Verbs like these are ever more often used in colloquial constructions like *Voz'mi s soboj poest' <popit'>* lit. ‘Take with you [something] to eat <to drink>’; cf. the observation that verbs are becoming more popular as hypernym descriptions of everyday items in Russian in (Iomdin B. 2010).

(Boguslavsky & Iomdin, 2010 a,b): *pit'* may be used as the name for what instantiates the second, object, valency of *pit'*. In any case, the meaning of consumption becomes grammatical. This is another instance of grammaticalization of this semantic field, supplementing the case of the verb *byt'* 'to be' in constructions like *Ja budu čaj* 'I will have tea' in contrast to **Ja budu knigu* 'I will [read?] the book' (reported elsewhere by L. Iomdin).

3.2 Criteria for disambiguation

3.2.1 Statistics

For *prosit'*, if the infinitive conveys the idea of consumption of food, in the vast majority of cases it is the subject infinitive. Out of 59 examples of *prosit' pit'* in the National Corpus of Russian, only one has an object infinitive: *Gospoda! u menja prošu pit' i est'* (N. S. Leskov) 'Gentlemen! I am asking you to eat and drink in my house'; out of 43 examples of *prosit' est'*, only 3 are subject infinitives.

3.2.2 Syntactic dependents of the head verb

Both *predlagat'* and *prosit'* have three semantic valencies: subject, object, and content. In the case of *predlagat'*, the object is expressed with a NP in dative, whereas for *prosit'* it can either be expressed by an NP in accusative (*prošu vas*) or by a NP in genitive with the preposition *u* (*prošu u vas*). For *prosit'*, this is a criterion for disambiguation: in *prosit' kogo-l. pit'* 'to ask someone to drink' *pit'* can only be interpreted as the object infinitive, while in *prosit' u kogo-l. pit'* 'to ask someone for a drink' *pit'* is definitely the subject infinitive.

The content in both verbs can be expressed either with an NP in accusative (*predlagat' <prosit'> jabloko* 'to offer <to ask for> an apple') or with an infinitive.

3.2.3 Syntactic dependents of the infinitive

For *prosit'*, the subject infinitive cannot govern any other words. Sentences like *Babuška prosit pit' čaj, Kto-to prosit sročno pit', Mama prosit est' pomedlennee* can only be viewed as containing object infinitives, meaning 'Grandmother is asking [everyone] to drink tea', 'Someone is asking [us] to drink immediately', 'Mom is asking [someone] to eat more slowly', respectively.

Acknowledgements

This research has been financed by a research program of History and Philology Branch of the Russian Academy of Sciences, a grant from the Russian Foundation of Basic Research (No. 11-06-00405), and a grant from the Russian Humanitarian Scientific Foundation (No. 10-04-00273). The authors are grateful to the Academy and both foundations. Valuable remarks by anonymous reviewers of the submitted version of the paper are also greatly appreciated.

Bibliography

Apresjan, Ju. 1974. *Leksičeskaja semantika. Sinonimičeskie sredstva jazyka*. Moscow: Nauka.

Apresjan, Ju., I. Boguslavsky, L. Iomdin & L. Tsinman. 2007. Lexical Functions in Actual NLP-Applications. In *Selected Lexical and Grammatical Issues in the Meaning–Text Theory. In honour of Igor Mel'čuk.* John Benjamins, Studies in Language Companion. Ser. 84. ISBN 978 90 272 3094 2. P. 199–230.

Apresjan, Ju. et al. 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. MTT 2003, First International Conference on Meaning–Text Theory (June 16–18 2003). Paris: École Normale Supérieure. P. 279–288.

Apresjan, Ju. et al. 2010. *Prospekt aktivnogo slovarja russkogo jazyka*. Moscow: Jazyki slavjanskix kul'tur.

Boguslavsky, I. & L. Iomdin. 2010. O valentnyx svojstvax odnogo širokogo klassa suščestvitel'nyx // Computational linguistics and intellectual technologies. Papers from the annual international conference "Dialogue" (2010). Moscow: RGGU. P. 47–54.

Boguslavsky, I. & L. Iomdin. 2010. Sintaksičeskie, semantičeskie i kommunikativnye svojstva avtoderivatov. In: *Gramatika i leksika u slovenskim jezicima*. Beograd: Matica srpska – Institut za srpski jezik. P. 15–17.

Boguslavsky, I. 2008. Aktantnoe povedenie adverbial'nyx derivatov. In *Dinamičeskie modeli: Slovo. Predloženie. Tekst.* Moscow: Jazyki slavjanskix kul'tur. P. 110–129.

Gardiner, A.H. 1932. *The Theory of Speech and Language*. Oxford: At the Clarendon Press, New York Oxford University Press, American Branch.

Iomdin, B. 2010. Russkaja bytovaja predmetnaja leksika: ontologija i opisanie. Proceedings of the 33rd Conference of young scientists and specialists of IITP RAS on informational technologies and systems (Gelendzhik, September 20–24 2010). http://www.itas-proceedings.iitp.ru/pdf/1569326461.pdf

Iomdin, B. in progress. *Davat'* and *dat'*. In Apresjan, Ju. et al. Aktivnyj slovar' russkogo jazyka. Vol. 1. To be published.

Uryson, E.V. 2004. *Mysl', ideja, duma*. In Apresjan, Ju. et al. *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka*. Ed. 2. Moscow-Vienna: Jazyki slavjanskoj kul'tury; Wiener Slawistischer Almanach. Sonderband 60. P. 551–556.

Illocutive Parenthetical Verbs in Russian

Lidija Iordanskaja and Igor Mel'čuk

OLST – Université de Montréal,
CP 6128 Centrte-ville, Montréal, Québec, H3C 3J7, Canada
lidija.iordanskaja@umontreal.ca | igor.melcuk@umontreal.ca

Abstract

The paper presents three Russian syntactic constructions in which can appear illocutive parenthetical verbal expressions [= IPVE], as in *The situation, **I believe**, is deteriorating*. The definitions of two relevant notions ('parenthetical' and 'illocutive') are proposed, the lexicographic presentation of verbs that take part in these constructions is discussed, and three semantic rules that ensure the production of correct Russian sentences with an IPVE are quoted.

Keywords

Russian syntax, parenthetical expressions, illocutive expressions, communicative structure of sentences, semantic rules, lexicographic description of some Russian parenthetical verbs.

1 Problem Stated

In this paper, we consider illocutive parenthetical verbal expressions [= IPVE] of the type in (1), where IVPEs are boldfaced:

(1) a. *The situation, **I think**, is deteriorating.*
b. *The situation, **the government hopes**, will improve.*
c. ***As 'The Times' reports**, the situation is deteriorating.*

Parenthetical expression E is called illocutive if, being part of the same sentence as clause *P*, E constitutes a comment by the Speaker semantically bearing on *P*: roughly, 'E–**sem**→P'. In other words, 'P' is an argument of 'E'—or, more precisely, of the main predicate of 'E'. E carries information about the clause *P* itself, i.e. its content or its form. In sharp contrast, a non-illocutive parenthetical expression carries additional information about the fact expressed by *P* rather than on *P* itself: for example, in the sentence *The situation—**such things already happened before**—will improve*, where E is boldfaced, 'P' (= 'the situation will improve') is not a semantic argument of 'E' (= 'such things happened before') .

We limit ourselves to illocutive parenthetical Es built around a finite verb form; such Es are known as "reduced clauses." Doing this, we leave out:

— Parenthetical expressions that are not illocutive ('P' is not an argument of 'E'): **(i)** Phatic parentheticals, used not to comment on *P* but to attract the Addressee's attention (*you know*; *Listen, …*) or to mark the Speaker's hesitation (...—*how could I put it?*—...).[1] **(ii)** Parenthetical clauses, as the boldfaced clause in (2), which present some additional secondary information about the state of affairs referred to by *P*:

(2) *The situation* ***(floods hit several new regions)*** *is deteriorating*.

— Expressions that are similar to parentheticals, without being parentheticals: **(i)** Direct Speech postposed introducers, as in *'Come here,'* ***whispered Helen***. Such an expression is the syntactic governor of Direct Speech (= *P*), while a parenthetical is a syntactic dependent of *P*. **(ii)** Autonomous main clauses whose syntactic link with *P* is asyndetic (= without an overt conjunction). Such a clause, unlike an IPVE, can carry sentential stress (Padučeva 1996: 322) ; for instance :

(3) a. Asyndetic coordination
The situation is deteriorating, | ***I knów***.

b. Asyndetic subordination
I knów*: the situation is deteriorating*.

Verbs in IPVEs have been called parenthetical [= $V_{(parenth)}$], beginning with Urmson 1952. The topic of parenthetical verbs is rather popular in linguistics: from the classic Urmson 1952 to Zalizniak and Padučeva [Z&P] 1987, Padučeva 1996, Schneider 2007, Blanche-Benveniste and Willems 2007 and Kahane and Pietrandrea [K&P] 2011. We deal here with parenthetical verbs in Russian; much of our data comes from Z&P 1987.

$V_{(parenth)}$s are specified by their semantic and syntactic properties.

The Semantic Properties of $V_{(parenth)}$s

1. A $V_{(parenth)}$ belongs to a vast semantic class of verbs denoting information processing in the human psyche and forming several subclasses: mental state verbs (THINK, BELIEVE, BE AFRAID, HOPE), mental activity verbs (DEMONSTRATE, DISCOVER), communication verbs (DECLARE) and perception verbs (SEE, HEAR).

2. An IPVE is necessarily a semantically positive statement (Apresjan 1978; Z&P 1987: 93-94). This means that:
— Either the meaning of a $V_{(parenth)}$ does not include an "internal" negation of the central component nor does $V_{(parenth)}$ have an "external" (= lexical) negation.
— Or the $V_{(parenth)}$ includes an internal negation but then it also has an external negation, which cancels the first one.

Thus, the Russian verbs SOMNEVAT´SJA 'doubt' ≈ '**not** be certain' and SKRYVAT´ 'hide' ≈ '**not** communicate…' are not $V_{(parenth)}$s. However, they can be used in an IPVE if supplied with an external negation:

(4) a. *Položenie,* ***ja ne somnevajus´****, uxudšaetsja* 'The situation, I don't doubt, is deteriorating'.

b. *Položenie,* ***ja ne skroju****, uxudšaetsja* 'The situation, I will not hide, is deteriorating'.

The Syntactic Property of $V_{(parenth)}$s

When used non-parenthetically, a $V_{(parenth)}$ takes, as a syntactic actant, a completive clause *P*: V–**synt**→*P*. But in a parenthetical use, for such a verb the syntactic dependency is inverted: V←**synt**–*P*.

---

[1] On "parasitic words" in Russian ("xmykan´e", "mekan´e" and "bljakan´e"), see Levontina and Shmelev 2007.

(5) a. ***Ja sčitaju,***→*čto položenie uxudšaetsja* 'I believe that the situation is deteriorating'.

b. *Položenie,* ***ja sčitaju,***←*uxudšaetsja* 'The situation, I believe, is deteriorating'.

The communication verbs OBSUŽDAT′ 'discuss' and PRIGLAŠAT′ 'invite' as well as the perception verb VOSPRINIMAT′ 'perceive' do not take a completive clause (**obsuždat′* ⟨**priglašat′, *vosprinimat′*⟩, *čto P*) and thus are not $V_{(parenth)}$s.

However, several Russian verbs have the three properties above, but nonetheless cannot be used in an IPVE: for instance, the belief verb ODOBRJAT′ 'approve'; the communication verb OB"JAVLJAT′ 'announce' (while ZAJAVLJAT′ 'declare' is a $V_{(parenth)}$); all verbs denoting a particular way of uttering: ŠEPTAT′ 'murmur', BORMOTAT′ 'mumble', KRIČAT′ 'shout', etc.; the perception verb OŠČUŠČAT′ '[to] sense' (but its quasi-synonym ČUVSTVOVAT′ 'feel' is a $V_{(parenth)}$).

> Our task in this paper is to propose a description of Russian $V_{(parenth)}$s and corresponding semantic rules that ensure the production of correct sentences with an IPVE.

Our research shows that $V_{(parenth)}$s have to be specified in the dictionary—that is, they must be described by individual dictionary rules rather than by some general rules. Such treatment is justified all the more since $V_{(parenth)}$s differ by the parenthetical constructions they can appear in. Thus, BOJAT′SJA**I.2** 'be afraid'[2] is used in an IPVE in the 1 sg of the present indicative, but not in the construction with the conjunction KAK 'as', while DOKAZAT′ 'demonstrate' manifests the inverse behavior; similarly, although BOJAT′SJA**I.2** refuses the IPVE with KAK, its quasi-synonym OPASAT′SJA 'fear' allows it:

(6) a. **Položenie, ja dokazyvaju, uxudšaetsja* 'The situation, I demonstrate, is deteriorating'. *vs.*
Položenie, kak ja dokazyvaju, uxudšaetsja 'The situation, as I demonstrate, is deteriorating'.

b. *Položenie, ja bojus′ ⟨ja opasajus′⟩, uxudšaetsja*
'The situation, I am afraid ⟨I fear⟩, is deteriorating'. *vs.*
*Položenie, kak ja opasajus′ ⟨*kak ja bojus⟩′, uxsudšaetsja*
'The situation, as I fear ⟨as I am afraid⟩, is deteriorating'.

2 The Communicative and Syntactic Roles of Illocutive Parenthetical Verbal Expressions

2.1 The Communicative Role of an IPVE

A common feature of all IPVEs at the semantic level is their communicative role: an IPVE is not part of the essential message expressed by the sentence in question, but indicates to the Addressee how he should interpret *P*. Urmson 1952: 496 compares an IPVE to a stage direction ("say it in a somber tone," etc.).

Formally, two communicative oppositions are relevant for an IPVE: Thematicity and Locutionality (Mel'čuk 2001: 93*ff*).

Thematicity. An IPVE does not belong to the communicative core of the sentence—to neither Rheme, nor Theme; it is within the communicative area of Specifiers (Mel'čuk 2001: 96*ff*). Specifiers are meanings that the Speaker uses in order to add information either on the situation P (e.g., a detached circumstantial of time or location, specifying temporal and spatial

[2] Lexicographic numbers for the verb BOJAT′SJA come from Iordanskaja and Mel'čuk 1990.

coordinates of P), or on the clause *P*; in this case, we speak of illocutive Specifiers. Since an IPVE constitutes a kind of a comment, by the Speaker, on the clause *P* (Bonami and Godard 2007: 259 and K&P 2011), it is an illocutive Specifier. The comment conveyed by an IPVE can bear on:

— The epistemological status of *P*, including the indication of the Speaker's source of the information « P » ('I believe', 'I swear it', 'declared the minister', 'as our calculations show').

— The subjective attitude of the Speaker or anybody else towards the content of *P* ('I am afraid', 'I regret', 'the government hopes').

— The linguistic form of *P* ('as say the Spaniards').

Locutionality. From the viewpoint of their locutionality, IPVEs are signalatives (Mel'čuk 2001: 245*ff*, 354*ff*)—more precisely, syntactic signalatives: their signalative character is expressed by a parenthetical syntactic construction. A signalative is a meaning 'σ' reflecting a psychological state of the Speaker or a rhetorical action by him such that he verbalizes it by signaling (rather than by communicating): a prototypical signalative expression does not allow for negation,[3] interrogation or free modification; it never constitutes an assertion in logical sense. Thus, the IPVEs in sentences (1) present a rhetorical action by the Speaker—namely, the introduction of an incidental comment concerning *P* (in this case, the signaled meaning is 'my source of the information «P» is E').

To make the notion of signalative clearer, here are examples of signalatives that are not IPVEs.

- First, there are lexical signalatives (= lexical units stored as such in the lexicon): interjections as *Wow!* or *Phew!*; parenthetical adverbials such as *unfortunately*, *to my sense* or *of course*; connector adverbials such as *in fact* or *for instance*; rhetorical conjunctions such as *although* or *since*; etc.
- Second, there are morphological signalatives, for instance, the imperative.
- Finally, there are syntactic signalatives, such as the Russian construction «V_{INF}-PARTICLE$_{\text{«TO»}}$ X V_{FIN}» (*Čitat'-**to** on čital, da...* lit. 'To.read-"to" he has.read, but…' ≈ 'Although he has read [it], but…'), which signals the skepticism of the Speaker with respect to X's action V.

At this point, two important remarks seem to be appropriate.

1. Note that an illocutive Specifier is not necessarily Signaled—that is, in our case, it does not to be expressed as an IPVE. Thus, in ***I believe** that the situation is deteriorating* the boldfaced matrix clause can be, in a particular context (for instance, as an answer to the question *What is happening there?*), an illocutive Specifier without being Signaled: it is not an IPVE.

2. The communicative status «Specifier + Signaled» of a meaning results in its weak communicative value; many researchers (e.g., Z&P 1987: 84 and K&P 2011) consider this property as essential for IPVEs. A weak communicative value also characterizes Backgrounded expressions, such as the boldfaced clause in *My friends* (***who live in Canada***) *like skiing*. But this is another communicative opposition, irrelevant in the context of this talk,—Perspective (Mel'čuk 2001: 198*ff*); an IPVE can be Neutral, as in (7a), or Backgrounded, as in (7b):

---

[3] More precisely, negation cannot bear on the **central** component of a signaled meaning. Thus, in the case of the imperative (which is a morphological signalative)—e.g., *Don't say this!*—the central component of an imperative meaning 'I want you to …' is not negated.

(7) a. *Položenie, **kak sčitajut vse**, uxudšaetsja*
'The situation, as everybody believes, is deteriorating'.
b. *Položenie **(kak sčitajut vse)** uxudšaetsja.*

Therefore, an IPVE cannot be defined simply as being communicatively Backgrounded.

The communicative role of an IPVE is manifested mainly by its prosody: an IPVE allows (or requires) pauses at its boundaries, carries—in a neutral context—a flat intonational contour and cannot have sentential stress (Z&P 1987: 81-82; Bonami and Godard 2007: 262 speak of "incidental prosody"). It is prosody that distinguishes a verbal IPVE in the initial position in the sentence (8a) from its non-parenthetical counterpart with the ellipsis of the conjunction ČTO 'that':

(8) a. (i) *Napominaju,* (|) *Maša bol´nÁ* '[I] remind, Masha is ill'. ≡
(ii) *Maša,* | *napominaju,* | *bol´nÁ* 'Masha, [I] remind, is ill'. ≡
b. (i) *NapominÁju,* || *Maša bol´nÁ* '[I] remind, Masha is ill'. ≡
(ii) *NapominÁju, čto Maša bol´nÁ* '[I] remind that Masha is ill'.

In (8a), *napominaju* '[I] remind' is an IPVE; in (8b), *napominaju* constitutes the matrix clause that governs its completive—asyndetically in (8b-i) and by means of ČTO (8b-ii).

2.2 The Syntactic Role of an IPVE

At the Deep-syntactic level, an IPVE depends on the head of the clause *P* by the deep-syntactic relation **APPEND**, which represents all kind of "extrastructural" constructions, manifesting weak subordination, such as sentence adverbs, parenthetical expressions, addresses, interjections, prolepses, etc. This type of subordination is opposed to strong subordination—that is, actants and modifiers/circumstantials.

3 Three Syntactic Constructions for IPVEs in Russian

The constructions under analysis will be illustrated by the verb SČITAT´ 'believe'.

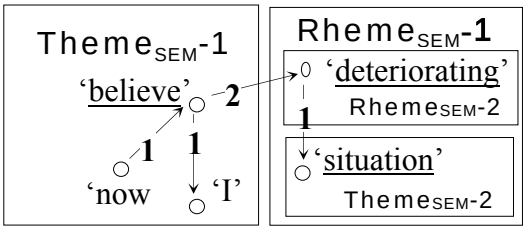
3.1 Non-Parenthetical Use of SČITAT´

Consider first a non-parenthetical use of the verb SČITAT´:

(9) a. *Ja sčitaju, čto položenie uxudšaetsja* 'I believe that the situation is deteriorating'.

Here is its semantic representation [= SemR]:

b.



In a communicative area, underlining indicates the communicatively dominant node, i.e., the semanteme that represents the minimal paraphrase of the area's meaning. The semanteme 'now' is an abbreviation that encodes the present indicative of the verb.

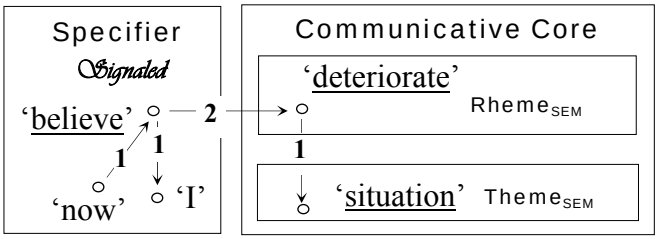
3.2 Parenthetical Uses of SČITAT′

Parenthetical-1 Construction

(10) a. ***Ja sčitaju**, položenie uxudšaetsja* 'I believe, the situation is deteriorating'. ≡
*Položenie, **ja sčitaju**, uxudšaetsja*. ≡ *Položenie uxudšaetsja, **ja sčitaju***.

The three sentences in (10a) have a common SemR, given in (10b):

b.



The communicative status of the signaled meaning 'I believe', realized by an IPVE, manifests itself in (10b) by the two properties introduced above:

— The meaning 'I believe' is a Specifier. Unlike the SemR in (9b), where this meaning belongs to the Theme, in (10b), it is not part of the communicative core.

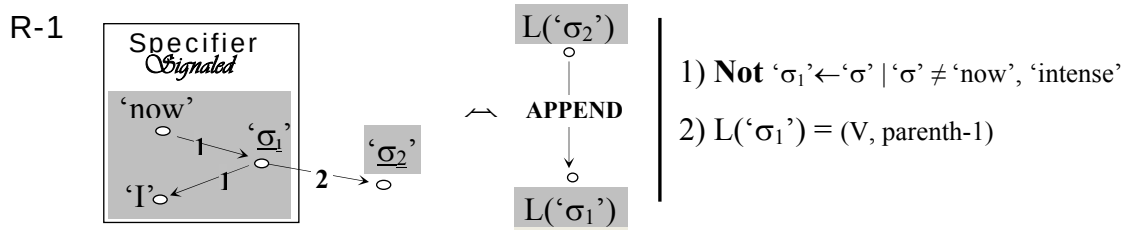
— The meaning 'I believe' is Signaled: the Speaker signals (rather than communicates) his epistemological attitude with respect to the clause *P*.

Properties of Parenthetical-1

1) Verb semantic class: mental state verbs (*ja sčitaju* 'I believe', *ja bojus′* 'I am afraid'), communication verbs (in a performative use: *ja nastaivaju* 'I insist', *ja garantiruju* 'I guarantee'), perception verbs (*ja slyšu* 'I hear'), but not mental activity verbs, such as *ja dokazyvaju* 'I demonstrate' or *ja zaključaju* 'I conclude'.
   A performative verb used in Parenthetical-1 construction signals the corresponding speech act rather than communicates it.
2) Syntactic subject: only JA 'I'; cf. **Položenie, **pravitel′stvo sčitaet**, uxudšaetsja* 'The situation, the government believes, is deteriorating'.
   The sentence *Položenie, **on sčitaet**, uxudšaetsja* manifests Parenthetical-2 construction.
3) Verb inflectional categories: the verb is in the active of the present indicative; cf. **Položenie, **ja sčital**, uxudšaetsja* lit. 'The situation, I believed, was deteriorating'.
   This sentence is possible, but then it represents Parenthetical-2.
4) Modification of the verb: no free modifiers; cf. **Položenie, **ja s nedavnix por sčitaju**, uxudšaetsja* 'The situation, I believe since recently, is deteriorating', except for a collocational intensifier: *Položenie, ja **tvërdo** sčitaju ⟨ja **točno** znaju⟩, uxudšaetsja* 'The situation, I strongly believe ⟨I know for a fact⟩, is deteriorating'.
5) Subordinability: no (**Ivan govorit, **čto** položenie, **ja sčitaju**, uxudšaetsja* 'Ivan says that …').
   Subordinability of an IPVE means the possibility of subordinating the whole sentence that contains this IVPE to a higher verb.
6) Omission of the subject: no (**Položenie, **sčitaju**, uxudšaetsja*).
7) Position of the subject: precedes the verb; cf. **Položenie, **sčitaju ja**, uxudšaetsja.*
   This sentence is possible, but then it represents Parenthetical-2.
8) Position of the IPVE: all three arrangements—before, inside and after *P*—are possible.
9) Prosody: two weak optional pauses on both sides (the second is a bit longer); low and flat contour; no sentential stress, no emphasis; low intensity.

Here and below, prosody is characterized quite approximately.

The transition between the SemR in (10b) and the DSyntS of (10a) is effected by semantic rule R-1. This rule is in fact a definition of the Parenthetical-1 construction. (Similarly, rules R-2 and R-3 define Parenthetical-2 and Parenthetical-3.)



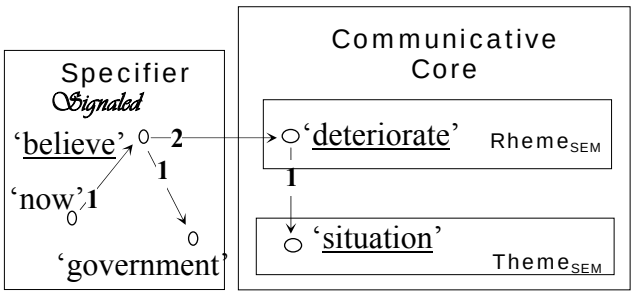
A shaded zone represents the context of the rule—that is, the elements that are not affected by the rule, but which control its application; the semantemes 'now' and 'I' are part of this context and are taken care of by corresponding rules. " | " indicates the conditions of the rule; L('σ') is the lexical expression of the meaning 'σ'. Condition 1 reflects the semantic constraints on the cooccurrence of the meaning 'σ1'.

Figure 1. Semantic Rule for Parenthetical-1 Construction

Parenthetical-2 Construction

(11) a. *Položenie,* ***sčitaet pravitel´stvo****, uxudšaetsja* 'The situation, believes the government, is deteriorating'. ≡ *Položenie uxudšaetsja,* ***sčitaet pravitel´stvo****.* ~ ****Sčitaet pravitel´stvo****, položenie uxudšaetsja.*

b.



Unlike Parenthetical-1, where the Experiencer of the signaled attitude towards P is necessarily the Speaker (= 'I'), Parenthetical-2 expresses an indication, by the Speaker, of the attitude of any person, including himself; as a result, the syntactic subject of the parenthetical verb in this construction can be of any grammatical person.

Properties of Parenthetical-2

1) Verb semantic class: mental state verbs (*sčitaet Ivan* 'believes Ivan', *nadeetsja Ivan* 'hopes Ivan'), mental activity verbs (*dokazyvaet Ivan* 'demonstrates Ivan', *uznaëm my* 'we learn'), communication verbs (*nastaivaet Ivan* 'insists Ivan') or perception verbs (*vidit Ivan* 'sees Ivan').
2) Syntactic subject: any nominal or pronominal expression; cf. *Položenie,* ***sčitaet pravitel´stvo****, uxudšaetsja* 'The situation, believes the government, is deteriorating'. This includes the 1sg: *Položenie,* ***sčitaju ja na osnove ètoj informacii****, uxudšaetsja* 'The situation, I believe based on this information, is deteriorating'.
3) Verb inflectional categories: voice, mood and tense are not constrained (except, of course, for the imperative); cf. *Položenie,* ***sčitalo***$_{\text{PAST}}$ ***pravitel´stvo****, uxudšalos´* lit. 'The situation, believed the government, was deteriorating'. ~ *Položenie,* ***budet sčitat´***$_{\text{FUTURE}}$ ***pravitel´stvo****, uxudšitsja.*
4) Modification of the verb: not constrained; cf. *Položenie,* ***sčitaet s nedavnix por pravitel´stvo****, uxudšaetsja* lit. 'The situation, believes since recently the government, is deteriorating'.

5) Subordinability: no (**Ivan govorit,* ***čto*** *položenie,* ***sčitaet pravitel´stvo***, *uxudšaetsja* 'Ivan says that …').

6) Omission of the subject: no (**Položenie,* ***sčitaet***, *uxudšaetsja*).

7) Position of the subject: a nominal subject follows the verb, cf. *Položenie,* ***sčitaet pravitel´stvo***, *uxudšaetsja*; a pronominal subject can follow or precede it, cf. *Položenie,* ***ja sčital*** *togda* ⟨***sčital ja*** *togda*⟩, *uxudšalos´*.

8) Position of the parenthetical: the initial position is impossible, cf. ****Sčitaet pravitel´stvo,*** *položenie uxudšaetsja.*

9) Prosody: two obligatory short pauses on both sides, the second being longer; a low and flat contour; no sentence accent; low intensity.

The transition between the SemR in (11b) and the DSyntS of (11a) is carried out by the semantic rule R-2:

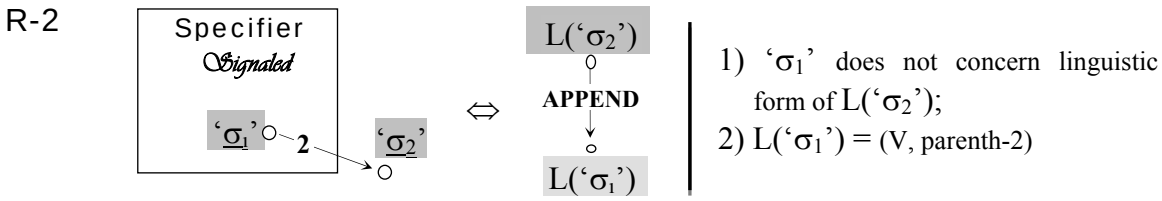


Figure 2. Semantic Rule for Parenthetical-2 Construction

Rule R-1 constitutes in fact a particular case of rule R-2; nevertheless, the introduction of two different parenthetical constructions is justified by at least the following two considerations:

— Certain verbs are used only in Parenthetical-1: e.g., BOJAT´SJA**I.2** 'be afraid' and NE SOMNEVAT´SJA 'not doubt' (see Section **5**).

— Parenthetical-1 is more constrained than Parenthetical-2—according to properties 2-4, even if Parenthetical-2 has a 1sg subject:

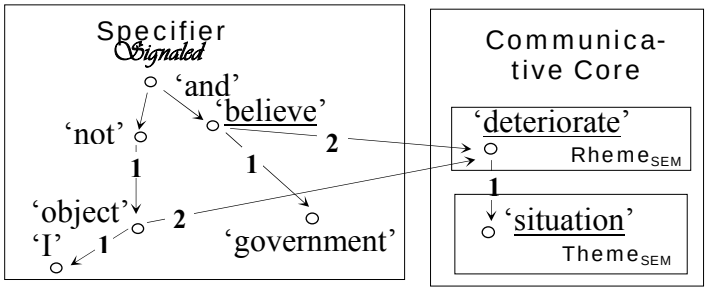
(12) a. *Položenie,* ***ja sčitaju*** *(*s nedavnix por), uxudšaetsja*
'The situation, I believe (since recently) is deteriorating'. *vs.*

b. *Položenie,* ***sčital ja v to vremja****, uxudšalos´*
'The situation, I believed then, was deteriorating'.

Parenthetical-3 Construction

(13) a. ***Kak sčitaet pravitel´stvo****, položenie uxudšaetsja*
'As believes the government, the situation is deteriorating'. ≡
Položenie, ***kak sčitaet pravitel´stvo***, *uxudšaetsja*. ≡
Položenie uxudšaetsja, ***kak sčitaet pravitel´stvo****.*

b.



Like Parenthetical-2, Parenthetical-3 expresses the attitude that, according to the Speaker, any person—including himself—has with respect to *P*. But Parenthetical-3 carries also an additional meaning: 'the Speaker does not have a contrary belief about [= does not object that] P' (Z&P 1987: 85-87); cf.:

(14) a. *Položenie,* ***sčitaet pravitel´stvo****, uxudšaetsja, no ja s ètim ne soglasen*
'The situation, believes the government, is deteriorating but I don't agree with this'. *vs.*

b. *Položenie,* ***kak sčitaet pravitel´stvo****, uxudšaetsja,* #*no ja s ètim ne soglasen*
'The situation, as believes the government, is deteriorating, but I don't agree with this'.

This additional meaning is a weak meaning: it is easily suppressed by a contradictory belief that the Speaker has stated before the appearance of the IPVE or after it, but then in a separate sentence:

(15) a. *Ja ne soglasen s tem, čto položenie,* ***kak sčitaet pravitel´stvo****, uxudšaetsja*
'I don't agree that the situation, as believes the government, is deteriorating'.

b. *Položenie ne uxudšaetsja,* ***kak sčitaet pravitel´stvo****, a naoborot, ulučšaetsja.*
'The situation is not deteriorating, as believes the government, but, on the contrary, is improving'.

c. *Položenie,* ***kak sčitaet pravitel´stvo****, uxudšaetsja. Po-moemu, odnako, ono ulučšaetsja*
'The situation, as believes the government, is deteriorating. In my opinion, however, it is improving'.

Unlike Parentheticals-1/2, Parenthetical-3 does not accept perception verbs:

(16) a. *Na kuxne, (ja) slyšu, kto-to xodit. ~ Na kuxne, slyšit Ivan, kto-to xodit*
'In the kitchen, I hear/hears Ivan, somebody is walking around'. *vs.*

b. **Na kuxne, kak ja slyšu, kto-to xodit. ~ *Na kuxne, kak slyšit Ivan, kto-to xodit.*
In the sentence *Ty, kak ja slyšal, polučil xorošee mesto* 'You, as I have heard, have landed a good position' we find a non-perceptional sense of the verb SLYŠAT´ (= 'learn by the way of speech').

Unlike Parenthetical-2, Parenthetical-3 can express a comment by the Speaker concerning the linguistic form of *P*:

(17) a. *Èto, kak govorjat v Odesse ⟨kak vyražajutsja odessity⟩, dve bol´šie raznicy*
'This, as they say in Odessa ⟨as express themselves the Odessites⟩, are two big differences'.

b. **Èto, govorjat v Odesse ⟨vyražajutsja odessity⟩, dve bol´šie raznicy.*

Properties of Parenthetical-3

1) Verb semantic class: mental state verbs (*kak sčitaet Ivan* 'as believes Ivan', *kak nadeetsja Ivan* 'as hopes Ivan'), mental activity verbs (*kak dokazyvaet Ivan* 'as demonstrates Ivan') and communication verbs (*kak zajavljaet Ivan* 'as declares Ivan'). A perception verb is possible only if its subject denotes a whole class of people (not specific individuals): *Ivan, kak vse videli ⟨*kak otec videl⟩, byl p´jan* 'Ivan, as everybody ⟨*Father⟩ saw, was drunk'.
2) Syntactic subject: can be any nominal or pronominal expression, cf. *Položenie,* ***kak sčitaet pravitel´stvo****, uxudšaetsja.*
3) Verb inflectional categories: voice, mood (except the imperative) and tense are not constrained; cf. *Položenie,* ***kak sčitalo/kak budet sčitat´/kak sčitalo by pravitel´stvo****, uxudšalos´.*
4) Modification: is not constrained; cf. *Položenie,* ***kak sčitaet s nedavnix por pravitel´stvo****, uxudšaetsja* lit. 'The situation, as believes since recently the government, is deteriorating'.
5) Subordinability: yes; cf. *Ivan govorit,* ***čto*** *položenie,* ***kak sčitaet pravitel´stvo****, uxudšaetsja*
'Ivan says that …'

6) Omission of the subject: no (**Položenie, **kak sčitaet**, uxudšaetsja*).

7) Position of the subject: not fixed (*Položenie, **kak pravitel´stvo sčitaet**, uxudšaetsja*).

8) Position of the parenthetical: all three arrangements are possible.

9) Prosody: two obligatory pauses on both sides, the second being longer; regular contour; no primary sentence stress, although secondary stress is possible (Z&P 1987: 87).

The transition "SemR (13b) ⇔ DSyntS of (13a)" is carried out by semantic rule R-3:

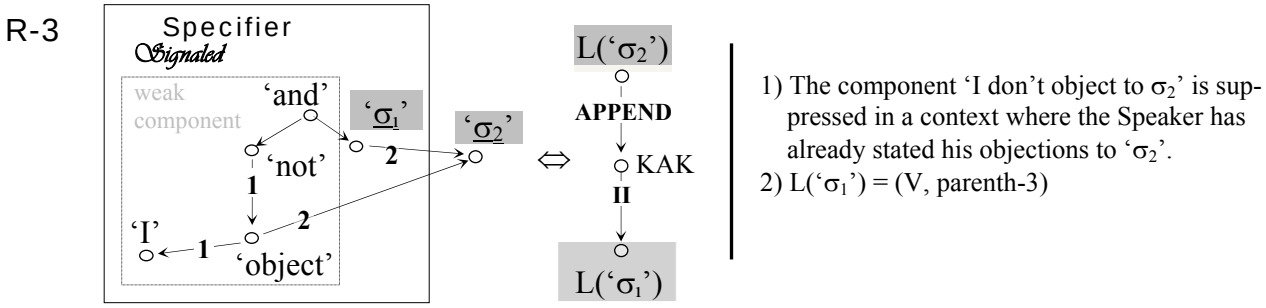


Figure 3. Semantic Rule for Parenthetical-3 Construction

4 Comparison of the Three Parenthetical Constructions

The three constructions share two defining features: each 1) expresses a Signaled Specifier and 2) syntactically depends on the head of the matrix clause *P* by the DSynt-relation **APPEND**.

At the same time, these constructions differ by several other features (see the corresponding lists above). In fact, Parenthetical-2 is freer (= has fewer constraints) than Parenthetical-1, and Parenthetical-3 is freer than Parenthetical-2. These differences correlate with the degree of the Speaker's involvement and with that of assertivity. On the one hand, the insubordinability of Parentheticals-1/2 is due to the fact that they can only express the comment by the Speaker himself; Parenthetical-3 is subordinable since it allows for the commentator to be a "substitute" of the Speaker—that is, the Observer or the Character in the narrative (Padučeva 2011). On the other hand, Parenthetical-1 does not constitute an assertion, Parenthetical-2 is closer to an assertion, and Parenthetical-3 is a quasi-assertion; Parenthetical-2/3 can be refuted by the Interlocutor, cf.:

(18) A: *Položenie, (kak) zajavilo včera pravitel´stvo, uxudšaetsja*
'The situation, (as) the government declared yesterday, is deteriorating'.
B: *Da net, ničego podobnogo pravitel´stvo ne zajavljalo*
'But no, the government did not declare anything like that'.

The "more assertive" character of Parentheticals-2/3 rules out their use in an interrogative sentence, which is possible for Parenthetical-1. This is explained by the fact that an assertion cannot follow an interrogation within the same sentence (Iordanskaja 1993: 173):

(19) *Položenie uxudšaetsja, ja bojus´? ~ *Položenie uxudšaetsja, boitsja pravitel´stvo? ~ *Položenie uxudšaetsja, kak boitsja pravitel´stvo?*

5 Parenthetical Verbs in the Dictionary

A Russian verb that has the semantic and syntactic properties licensing its participation in an IVPE is not necessarily usable in one: see the end of Section 1. Moreover, those verbs that do participate in IVPEs differ in the type of construction they can be used in—each $V_{(parenth)}$ is characterized by the constructions it accepts. Therefore:

> The dictionary article of a $V_{(parenth)}$ must indicate the type of construction this V can be used in: «parenth-1», «parenth-2», «parenth-3», or a combination thereof.

The question arises: Is such a method necessary and sufficient?

Necessity has been demonstrated above; let us add a few more examples.

— One verb may be a $V_{(parenth)}$ while a semantically and syntactically similar one is not: POLAGAT′ 'suppose' is a $V_{(parenth)}$, but PRINIMAT′ 'accept' is not; the same is true about UBEŽDAT′ 'convince' (in construction with PYTAT′SJA 'try') *vs.* UGOVARIVAT′ 'persuade' (even with PYTAT′SJA); contrary to the English $V_{(parenth)}$ REGRET, its Russian equivalent SOŽALET′ 'regret' is not a $V_{(parenth)}$, although the corresponding meaning can be expressed in Russian (by the adverbial K SOŽALENIJU 'regrettably').

— Both verbs are $V_{(parenth)}$s, but do not accept the same constructions: the verb VERIT′ 'believe = have faith' and the verbal expressions BYT′ UVEREN 'be sure' and BYT′ SOGLASEN 'agree' cannot be used in Parentheticals-2/3, but SČITAT′ 'believe' can. The verbs DOKAZYVAT′ 'prove' and POKAZYVAT′ 'demonstrate' are excluded from Parenthetical-1, but are usable in Parentheticals-2/3; and UTVERŽDAT′ 'affirm' participates in all three constructions.

Sufficiency. The proposed marking of parenthetical verbs is not sufficient: it has to be supplemented by finer individual features.

Thus:

— The $V_{(parenth-1)}$ SKRYVAT′ 'hide' can be used, unlike typical $V_{(parenth-1)}$s, in the future and allows for the omission of the subject:

(20) *Položenie, **(ja) ne skroju** ⟨= **ne stanu skryvat′**⟩, uxudšaetsja*
'The situation, I won't hide, is deteriorating'.

— The $V_{(parenth-1)}$ BOJAT′SJA**I.1b** ≈ 'be afraid' can be used in Parenthetical-3, but only in the past and with the emphatic particle I (Z&P 1987: 95):

(21) a. *Položenie, **kak ja ⟨on⟩ i bojalsja**, uxudšilos′* 'The situation, as I ⟨he⟩ feared, deteriorated'.

b. **Položenie, kak ja ⟨on⟩ bojalsja, uxudšilos′* 'The situation, as I ⟨he⟩ feared, deteriorated'.

— The $V_{(parenth-1)}$ BOJAT′SJA**I.2** 'be afraid', but not SČITAT′ 'believe', allows for the omission of the subject:

(22) *Položenie, **bojus′** ⟨*sčitaju⟩, uxudšaetsja*
lit. 'The situation, am afraid ⟨*believe⟩, is deteriorating'.

— NADEJAT′SJA 'hope' allows the 1pl of the imperative, but OPASAT′SJA 'fear' does not:

(23) *Položenie, **budem nadejat′sja**, ulučšitsja* 'The situation, let's hope, will improve'. *vs.*
Položenie, **budem opasat′sja, uxudšitsja* 'The situation, let's fear, will deteriorate'.

— ZNAT′ 'know' can appear in Parentheticals-2/3, but only if its subject denotes a whole class of people (rather than an individual):

(24) a. *Položenie, **(kak) znaet každyj durak**, uxudšaetsja*
'The situation, as any fool knows, is deteriorating'. *vs.*

b. **Položenie**, **(kak) znaet Ivan**, uxudšaetsja.*

In Parenthetical-3, the subject of ZNAT′ can also be the Addressee:

c. *Položenie, **kak vy znaete**, uxudšaetsja* 'The situation, as you know, is deteriorating'.

To sum up: any $V_{(parenth)}$ has to be specified in the dictionary by means of syntactic features « parenth-1 », « parenth-2 » and « parenth-3 », supplemented with additional individual features and conditions under which such a use is possible. For instance:

SČITAT′	: (parenth-1, parenth-2, parenth-3)	BYT′ UVEREN	: (parenth-1, parenth-2)
BOJAT′SJA**I.1b**	: (parenth-2; parenth-3 | particle I, in the past)	VIDET′	: (parenth-1, parenth-2)
BOJAT′SJA**I.2**	: (parenth-1, the subject is omissible)	UKAZYVAT′	: (parenth-2, parenth-3)
SOMNEVAT′SJA	: (parenth-1 | avec NE 'not')	DOKAZYVAT′	: (parenth-3)
NADEJAT′SJA**3**	: (parenth-1, the subject is omissible, +1 pl imperative)	ZNAT′	: (parenth-1; parenth-2/3 |subject denotes a class of people)

6 Conclusion

To close our discussion of Russian parenthetical verbs, we would like to make the following five remarks.

1. Our study allows us to formulate a definition of parenthetical expression that covers all verbal parentheticals (quoted at the beginning of this paper) and adverbials such as *unfortunately*, *according to John, frankly, as they say*, etc.

Definition 1: Parenthetical Expression

> An expression E linked to the clause *P* within a sentence is called parenthetical, if and only if:
> 1) in the semantic-communicative structure of the sentence, 'E' is a Signaled Specifier;
> 2) in the deep-syntactic structure, E depends, by the syntactic relation **APPEND**, on the head of *P*.

Let us emphasize that 'being parenthetical' is a **syntactic** property of an expression that reflects its semantic-communicative particularities. From the **semantic** viewpoint, parenthetical expressions are subdivided into illocutive ones (which take the rest of the sentence as a semantic argument) and non-illocutive (where this is not the case). The definition of illocutive parenthetical expressions is now straightforward.

Definition 2: Illocutive Parenthetical Expression

> A parenthetical expression E is called illocutive, if and only if, E semantically bears on the clause *P*: 'E–**sem**→P'.

We propose a broader sense of the term *illocutive*, as compared to its definition in Iordanskaja 1993: there, an expression E is called illocutive only if it semantically bears on the **fact of uttering** of *P*; here, to be illocutive, E has to bear on *P*, covering both **its uttering** and **its content/form**. We by no means insist on this terminological solution; perhaps a better way would be to think of a different term.

2. IVPEs represent a case of syntactic signalatives, which exist along the well-known lexical signalatives (interjections, textual connectors, etc.) and morphological (e.g., the imperative) signalatives.

3. Russian $V_{(parenth)}$s have to be specified (in the dictionary) by syntactic features « parenth-1/2/3 » and some additional features; it seems impossible to give a reliable semantic characterization of the class of verbs participating in the same parenthetical construction.

4. However, some local (= partial) generalizations are possible; for instance:

— No parenthetical construction allows a speech verb whose meaning includes the manner of speaking: **Položenie, (kak) bormočet on, uxudšaetsja* 'The situation, (as) he mumbles, is deteriorating'.

— Parenthetical-1 does not allow verbs of mental activity: **Polozenie, ja obnaruživaju ⟨dokazyvaju, vyjasnjaju⟩, uxudšaetsja* 'The situation, I discover ⟨demonstrate, establish⟩, is deteriorating'.

— Parenthetical-3 allows neither a verb of perception (except the case of a "general-indefinite" subject), nor a verb of belief whose meaning includes the manner of believing (Z&P

1987: 87; **V kuxne, kak on slyšal, kto-to xodil* 'In the kitchen, as he heard, somebody was walking around', **Položenie, kak on ne somnevaetsja ⟨uveren, gotov pokljast´sja⟩, uxudšaetsja* 'The situation, as he does not doubt ⟨is sure, is ready to swear⟩, is deteriorating'.

5. Russian has a further type of IVPE, which we did not discuss here: with monoargumential verbs and verbal expressions, of the type KAZAT´SJA 'seem' or STAT´ IZVESTNYM 'become known':

(25) *Položenie, mne kažetsja ⟨kak Ivanu stalo izvestno včera⟩, uxudšaetsja*
'The situation, it seems to me ⟨as it became known to Ivan yesterday⟩, is deteriorating'.

Here, 'P' is the semantic actant **1** of 'E' (rather than **2**, as in our case). Therefore, additional semantic rules are required, which, however, do not pose any theoretical difficulty.

Acknowledgments

The impetus for this paper came from the reading the K&P 2011 manuscript, for which we cordially thank the authors. The text has been read and commented by D. Beck, I. Boguslavskij, S. Kahane, R. Laskowski, J. Milićević, E. Padučeva and E. Savvina; to all of them we express our most heartfelt gratitude.

References

Apresjan, Ju. 1978[1995]. Jazykovaja anomalija i logičeskoe protivorečie. In Apresjan, Ju. *Izbrannye trudy. Tom II*. Moskva: Jazyki russkoj kul´tury, 598-621.

Blanche-Benveniste, C. and Willems, D. 2007. Un nouveau regard sur les verbes « faibles ». *Bulletin de la Société de Linguistique de Paris*, 102(1):217-254.

Bonami, O. and Godard, D. 2007. Quelle syntaxe, incidemment, pour les adverbs incidents ? *Bulletin de la Société de Linguistique de Paris*, 102(1):255-284.

Iordanskaja, L. 1993. Pour une description lexicographique des conjonctions du français contemporain. *Le Français Moderne*, 61(2):159-190.

Iordanskaja, L. and Mel'čuk, I. 1990. Semantics of Two Emotion Verbs in Russian: BOJAT´SJA '(to) be afraid' and NADEJAT´SJA '(to) hope'. *Australian Journal of Linguistics*, 10(2): 307-357. See also Mel'čuk, I. 1995. *The Russian Language in the Meaning-Text Perspective*. Moskva/ Wien: Jazyki russkoj kul´tury/Wiener Slawistischer Almanach, 81-124.

Kahane, S. and Pietrandrea, P. 2011. Les parenthétiques comme « Unités Illocutoires Associées » : une approche macrosyntaxique. In Avanzi, M. and Glikman, J. (eds.). *Les Verbes Parenthétiques : Hypotaxe, Parataxe ou Parenthèse ?, Linx.*

Levontina, I. and Shmelev, A. 2007. False Emptiness: Are So-called "Parasitical Words" Really Semantically Void? In Gerdes, K., Reuther, T. and Wanner, L. (eds.), *Meaning-Text Theory 2007*, Wiener Slawistischer Almanach, München-Wien, 259-268.

Mel'čuk, I. 2001. *Communicative Organization in Natural Language. The Semantic-Communicative Structure of Sentences*, Amsterdam/Philadelphia: Benjamins.

Padučeva, E. 1996. *Semantičeskie issledovanija*. Moskva: Jazyki russkoj kul´tury.

Padučeva, E. 2006. Vvodnye glagoly: rečevoj i narrativnyj režim interpretacii. In Moldovan, A. (ed.), *Sbornik statej k 60-letiju V.M. Živova*, Moskva: Jazyki slavjanskix kul´tur, 607-623.

Padučeva, E. 2011. Ègocentričeskie valentnosti i dekonstrukcija govorjaščego. *Voprosy jazykoznanija*, No. 3.

Schneider, S. 2009. *Reduced Parenthetical Clauses as Mitigators*. Amsterdam/Philadelphia: Benjamins.

Urmson, J. 1952. Parenthetical Verbs. *Mind*, 61(244):480-496.

Zaliznjak, Anna and Padučeva, E. 1987. O semantike vvodnogo upotreblenija glagolov. In Eršov, A. (ed.), *Vorposy kibernetiki. Prikladnye aspekty lingvističeskoj teorii*, Moskva: AN SSSR, 80-96. See also: Zaliznjak, Anna. 2006. *Mnogoznačnost´ v jazyke i sposoby eë predstavlenija*. Moskva: Jazyki slavjanskix kul´tur, 463-477; Padučeva 1996: 321-334.

Presenting collocates in a dictionary of computing and the Internet according to user needs

Anne-Laure Jousse (1), Marie-Claude L'Homme (1),
Patrick Leroyer (2) and Benoît Robichaud (1)

(1) Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
anne-laure.jousse|mc.lhomme|benoit.robichaud@umontreal.ca

(2) Center for lexicography
Aarhus University
pl@asb.dk

Abstract

This paper presents a novel method for organizing and presenting collocations in a specialized dictionary of computing and the Internet. This work is undertaken in order to meet a specific user need, i.e. that of searching for a collocate (or a short list of collocates) that expresses a specific meaning in a text production situation. The model we suggest is based on lexical functions (=LFs) that formally encode syntactic, argumental and semantic properties of collocations. LFs are grouped in larger semantic classes (e.g., USE, CREATE, PLACE SOMEWHERE, etc.). The model is in the process of being implemented in the online version of the dictionary. Users are prompted with generic meanings associated with the classes we created and they can then select verb and noun collocates that express more specific meanings. The article describes the model for grouping collocations and its implementation. Finally we present a small pilot study that was conducted in order to gather some user feedback on the usefulness of the method.

Keywords

Collocations, lexical functions, specialized dictionary, text production, computing, Internet.

1 Providing onomasiological access to collocates

An increasing number of dictionaries (general and specialized) present collocations within their entries. Some of these dictionaries are online (DAFLES, DiCE, DiCoInfo) and thus can provide different access paths to users based on how collocations were encoded in the first place. Depending on the dictionary use situation, users will generally consider the presentation of collocations as extremely useful information. In specialized language, collocations are essential to the production of specialized discourse in accordance with the writing and genre conventions used by professionals. Even if compilers of dictionaries may take it for granted that the issue of selecting the collocations is settled, they still face a number of challenges, among which are the following:

1. How should collocations be presented in specific entries, especially in online dictionaries in which high numbers of collocations are listed?

2. Should a description of their meaning accompany collocations? What should this description look like?

3. What should be done to ensure the functional usability of data presentation and access?

This article will attempt to answer all three questions focusing on a specific user need, namely that of a user searching for a collocate (or a short list of collocates) that expresses a specific meaning in a text production situation. In other words, the user knows which meaning should be expressed but does not know the specific, conventionalized wording itself. More specifically, our aim is to design an access method that would allow users consulting an online dictionary of computing and the Internet to obtain answers to the questions such as those listed below:

- Which verbs express the idea of "using" a dialog box (utiliser une boîte de dialogue)?[1] Answer: *activer, afficher, ouvrir une boîte de dialogue* (enable, display, open a dialog box)
- Which verbs express the typical activities carried out by a programmer? Answer: *le programmeur écrit …, débogue .., développe …, corrige, programme* (the programmer writes …, debugs …, develops …, programs)

The remainder of the article is organized as follows. Section 2 presents previous attempts at organizing and presenting collocations in dictionaries. Section 3 describes the DiCoInfo and gives more specific details about how collocations are encoded in this dictionary. Section 4 explains how collocations were grouped and how the new access method was implemented. Finally, Section 5 gives the details of a pilot study that was carried out in order to check if the model suited the targeted user needs. Section 6 briefly outlines future work.

2 Previous work

In recent years, a few (printed or electronic) dictionaries and lexical databases have attempted to present collocations according to one or several organizing principles. In printed dictionaries (cf. Tutin, 2010) or lexical databases, two main organizing methods are preferred. Most reference works organize collocations syntactically (LTP Dictionary of Selected Collocations [Hill & Lewis, 2002], Antidote [Charest *et al*., 2007]); others add a semantic layer to the syntactic classification that is seldom made explicit and that merely presents itself as lists of synonymous collocates (Le Robert [Le Fur, 2007], BBI Dictionary of English Word Combinations [Benson *et al*., 1997], Oxford Collocations Dictionary for Students of English [McIntosh, C., 2009]).

More sophisticated repositories combine several organization principles (semantic, morphological, and syntactic) (Base Lexicale du Français [Verlinde et al., 2006]), le *Dictionnaire d'apprentissage du français des affaires* ([Binon et al., 2009]), Elexiko [Klosa et al., 2006]). However, when a proper modeling of semantic relationships between lexical units and collocates is lacking, a complete automatic grouping of collocates in semantic categories can simply not be considered. This leads lexicographers to classify relationships in an ad hoc manner and often to resort to introspective methods (Cinkova and Hanks, 2010).

If an overall organization of collocations is considered from the point of view of the entire lexicon of a language, it can only be carried out based on a solid formalization of lexical

[1] The implementation of the method is currently carried out in the French version of the dictionary. However, an extension to the English and Spanish versions will be done shortly.

relationships. Indeed, this formalization allows for a generalization of the classification to larger parts of the lexicon (based on the development of a model on a representative sample): the vocabulary associated with a specialized subject field, the lexicon of a specific language or a group of languages. Formalization is also necessary for processing and manipulating data.

Hence, the formal system of Lexical Functions (=LFs) (presented in Section 3.3) used in Explanatory Combinatorial Lexicology (=ECL) is perfectly adapted for this kind of task. Various lexical resources based on this framework have been developed during the past few decades. Some take the form of lexical databases (DiCo [Polguère, 2000], DiCE [Alonso Ramos, 2004], DiCoInfo [L'Homme, 2011]); others are tools for language learning (Callex [Diachenko, 2006], Callex-Esp [Boguslavsky et al., 2006]). In all these, LFs are used to represent syntactic and semantic properties of lexical relationships.

A more specific proposal was made for representing relationships in a lexical database (Jousse, 2007; Jousse, 2010; Jousse et al. 2008). The authors use the existing database DiCo and suggest that paradigmatic as well as syntagmatic relationships be organized in such as way that users can follow different paths for browsing parts of the lexicon. The method takes into account: 1. a semantic organization that allows users to access relationships onomasiologically; 2. a syntactic organization for selecting a collocation based on a specific syntactic configuration; 3. a classification based on parts of speech; and, finally, 4. an organization taking into account communicative criteria that will suggest collocates according to the argument that is highlighted in a specific collocation (for example, in *X gives a call to Y*, X is highlighted, whereas Y is emphasized in *Y receive a call from X*). This method, however, has not been implemented yet.

The DiCoInfo, presented in the next section, is based in part on the same theoretical principles as the DiCo and thus lends itself to a similar formal organization of its collocations.

3 The DiCoInfo: form and functions

3.1 The DiCoInfo

The DiCoInfo, *Dictionnaire fondamental de l'informatique et de l'Internet* is an online dictionary (http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi) that provides information on terms pertaining the fields of computing and the Internet (e.g., *access, configure, dynamic, read, software*). Currently, the DiCoInfo contains over 1,000 articles in French and approx. 700 articles in English (a Spanish version is also under development).

The methodology for compiling the DiCoInfo is based on a combination of automated and manual methods. A series of steps (selection of terms, collection of example sentences, writing of entries) is carried out by terminologists more or less in the same order. The data is encoded in an XML structure and then converted into HTML pages for the purpose of publishing its content on the Internet.

3.2 Recent improvements for functional purposes

Work on a more adaptive and user-oriented access to data in the DiCoInfo was initiated back in 2009. It paved the way for the development of automatic access to translations of collocations (L'Homme & Leroyer, 2009). The software application was adapted a year later, and now includes a new version of the search engine, enabling user-friendly, automatic access to translations of collocations in French, English, and Spanish. Collocations sharing identical

semantic and syntactic properties were linked up by means of the encoding of LFs (L'Homme, Leroyer & Robichaud, 2010).

Providing onomasiological access to collocations was also considered. This could be done in part by using the paraphrasing of lexical relations in the database (cf. Section 3.3). The idea was to provide assistance in text production situations (in L1 or in L2) in which the user knows the meaning of a phraseological unit but is searching for the appropriate collocates that appear in combination with that unit. However, at the time, this new type of access was devised but not concretely designed or implemented.

3.3 Lexical functions in the DiCoInfo

One of the most important data categories in the DiCoInfo is that of lexical relationships. Each entry contains a list of lexical units sharing with the head word a paradigmatic or a syntagmatic relationship (synonymy, antonymy, syntactic derivation, collocates, etc.). All lexically-related units are explained using two different systems: 1. the system of LFs (Mel'čuk et al., 1995, Mel'čuk et al., 1984-1999); 2. a less formal and language-dependent explanation designed to be more transparent for users (these explanations are based on a proposal made by Polguère, 2003). Table 1 gives a few examples of how the collocations are explained in the database.[2]

Key word	Collocate	Lexical function	Explanation
programmer	the ~ write …	Fact2	The p. acts on the program
dialog box	open a ~	Real1	The user uses a d.
program	quit a ~	FinReal1	The user stops using a p.
Internet	browse the ~	Real1	The user uses the I.
keyboard	enter … on a ~	Labreal12	The user uses a k. to act on the data
account	access an ~	IncepReal1	The user starts using an a.

Table 1: Collocations, lexical functions and explanations

In the online version of the dictionary, lexical relationships (among which collocations) were all listed in the form of a table. Collocations were presented in a section called "Combinations" that was very long and difficult to read in some entries. For example, in the article devoted to "fichier" (file), approximately 100 collocates were listed.

4 A model for grouping and browsing collocations

4.1 Grouping collocations in transparent classes

In the DiCoInfo, LFs were first grouped into more general semantic classes to allow users to access collocations onomasiologically (from the meaning to the collocate). We analyzed the relationships that had been encoded in the DiCoInfo and found that specific classes were dominant (for example, since the field of computing needs terms that denote entities, many collocates express the idea of USE or MAKE STH WORK). The semantic classes were defined based on the results of the analysis of corpus data in order to ensure that they would capture recurrent relationships in a balanced way. When defining the classes, terminologists started using some frequent collocational relationships. As was said above, LFs encoding USE and TO MAKE STH WORK ($Real_i$, $Labreal_{ij}$), etc. were particularly productive. All LFs

[2] Wherever possible, each LF is explained with a unique gloss. As far as the phrasing of glosses is concerned, we aim to provide – as much as possible – a transparent and natural explanation. In addition, the phrasing may vary slightly according to the base of the collocation.

encoding a typical use were first grouped regardless of arguments and secondary meanings involved ("Use", "Use for something", "Agent uses", "Other argument or external participant uses") into intermediate classes. Then, more generic classed were defined. While conducting this analysis, we also took into consideration that the user might need to access specific pieces of information concerning collocates.

Lexical function	Intermediate class	Generic class
Caus1Able1Fact0, Caus1Able1Real1, Caus1Able1Real3, PermFact0, Perm1Fact0	**Permettre l'utilisation / Activer**	**UTILISER / NE PAS UTILISER**
Prepar@, Prepar1, Prepar1Fact0, Prepar1Real1, Prepar2Real3, Prepar@Fact0, De_nouveauPrepar1, De_nouveauPrepar1Fact0	**Préparer l'utilisation / le fonctionnement**	
IncepLabreal12, IncepReal@, IncepReal1, IncepReal2, IncepReal3	**Commencer à utiliser / Apparaître**	
Caus1Fact0, Caus@Fact0, Labreal@2, Labreal12, Labreal123, QLabreal12, Real@, Real1, Real12, Real123	**Utiliser / Faire fonctionner**	
FinLabreal12, FinReal1, FinReal2, Liqu1Fact0, Liqu@Fact0	**Cesser d'utiliser / de faire fonctionner**	

Table 2: Grouping of LFs under intermediate and generic classes

CRÉER/SUPPRIMER **Créer / Faire apparaître** *créer, générer un fichier*	**TO CREATE/TO DELETE** **To create or display** *to create, to generate a file*
Supprimer / Détruire *supprimer, effacer un fichier*	**To delete / eliminate** *to delete a file*
TRANSFORMER **Transformer** *crypter, convertir un fichier*	**TO TRANSFORM** **To transform** *to encrypt, to convert a file*
Diminuer / Réduire *comprimer un fichier*	**To reduce** *to compress a file*
UTILISER/NE PAS UTILISER **Préparer l'utilisation / Le fonctionnement** *installer, rechercher un fichier*	**TO USE/ USE NOT** **To prepare for use, operation** *to install a file, to search for a file*
Commencer à utiliser / Apparaître *charger, ouvrir un fichier*	**To start to use / to appear** *lo load, to open a file*
Utiliser / Faire fonctionner *traiter, éditer un fichier*	**To use/to make sth work** *to process, to edit a file*
Cesser d'utiliser / De faire fonctionner *fermer un fichier*	**To stop using/working** *to close a file*
METTRE QUELQUE PART **Ajouter à / Mettre dans** *joindre un fichier à un courriel*	**TO PLACE SOMEWHERE** **To add, to place in** *to attach a file to an e-mail*
Stocker ~ quelque part *archiver, télécharger un fichier*	**To store somewhere** *to archive, to download a file*
Transférer *exporter, transférer un fichier*	**To transfer** *to export, to forward a file*
Extraire / Sortir de *désarchiver un fichier*	**To extract/Quit** *to extract a file*
IDENTIFIER **Identifier** *nommer un fichier*	**TO IDENTIFY** **To identify** *to name a file*

Table 3: Classification examples for a subset of collocations of 'fichier'

Classes were defined according to the main relationships that can be observed in the field of computing and not according to general principles that could apply to general language (as in Jousse, 2010). Table 2 shows how we grouped recurrent LFs in intermediate classes under the generic class of **UTILISER/NE PAS UTILISER** (TO USE/USE NOT).

In fact, some of the classes identified could apply to other subject fields and to general language, but we focused on what could be observed in our database. We also believe that our method for grouping and balancing them is closely related to the subject field we are dealing with. Hence, it is most likely that some classes might not have the same value in other specialized subject fields. Examples in point are **METTRE QUELQUE PART** (TO PLACE SOMEWHERE, used to capture collocates such as *store* and *save*) and **TRANSFORMER** (TO TRANSFORM, used to capture verbs like *format* and *compile*).

We also tried to limit the number of different classes as much as possible in order not to overload the interface with long lists of classes and help users memorize them more easily. Up to now, 9 generic classes and 45 intermediate classes have been defined. Intermediate classes contain approx. 300 different LFs. Table 3 gives examples of classes that are displaying when looking at the entry "fichier" (file).

It is worth pointing out that some collocates have complex meanings and can be classified into more than one class. This is the case with *exporter* (export) that conveys both the meanings of transforming and transferring. We thus placed the verb in two classes (**TRANSFORMER** and **TRANSFÉRER**) thinking that users might access collocates from these different access points.

4.2 Browsing collocations in the dictionary

As mentioned previously, the DiCoInfo is an XML-based dictionary. The classes are naturally declared and organized in a hierarchy that is also modeled as an XML structure suited for such interlocking. At this first stage of the implementation, the hierarchy is limited to four distinct types of classes strictly ordered: a *root*, the generic and intermediate classes (such as **CRÉER/SUPPRIMER**: TO CREATE/TO DELETE; and **Supprimer/Détruire**: To delete / Eliminate), and lastly terminal classes that are the LFs names. Yet, this simple implementation has an essential feature: it allows intermediate and terminal classes to have more than one parent. This characteristic contributes significantly to improve browsing paths as it makes it *a priori* possible to describe parallel access paths to collocations based on different *points of view* (as argued in Section 2); or to classify more accurately collocations that have complex meanings (as exemplified with the *exporter* case in Section 4.1). Figure 1 bellow shows the low complexity of the actual hierarchy and the respective proportion of the different classes (terminal classes are not connected to intermediate ones for clarity).

To ease their management, the inventory and organization of collocation classes that emerge from the grouping analysis are stated as data independently of the dictionary entries (i.e. with the exception of LFs that formalized the links between head words and collocates, no reference to other classes is made within the entries) and the programs that manipulate them. This way, terminologists may easily access and modify at will the organization of the classes to shape the browsing paths faster without having to edit the entries or the programs. The online version of the DiCoInfo (including pages from the search interface) are created by means of XSL transformations of the initial XML dictionary files into HTML pages. While previous versions of the program that generates the pages simply listed the collocations of a dictionary entry in one long table, the present version loads the class hierarchy as an additional data structure along with the dictionary files, and then displays entries within an

outline view (or *tree view*) section that holds the collocations according to the hierarchy. This new section is first presented as an ordinary hyperlink. By clicking on this hyperlink, users open the hierarchy and may select different *branches* (or *nodes*) according to the class names presented and their search needs (as shown in Figure 2). Ultimately, browsing paths reach the terminal classes and short tables are presented with the usual information about collocations.

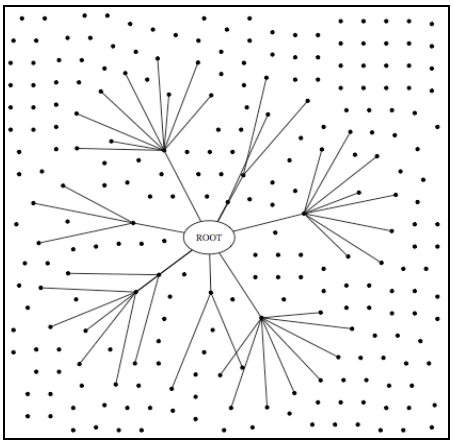


Figure 1: Spring view of the class hierarchy

Figure 2: Outline view of the class hierarchy

5 Usability test

5.1 Background and design

One of the main requirements for the application developed in the framework of this research project is to adapt data access and presentation in order to cater more efficiently for the specific needs of intended users. Hence, we decided to conduct a small-scale pilot-study of the usability of the new data presentation. As explained in more detail in section 3.2, the new presentation is aimed at providing specific assistance in text production situations in L1 or in L2, in which users know the meaning of a collocation but do not recall its components (often verbs with a specialized meaning). Access is then provided to help users find potential candidates for the relevant collocates. This section will attempt to answer the last question asked at the beginning of this paper: Can we assess the usability of the new presentation? By asking two more specific questions: Does it provide a fast and easy access? Can we assess its overall efficiency?

In order to answer these questions, we decided to obtain feedback from the users of the DiCoInfo by means of a test in a controlled experimental environment. The participants were selected from among the intended core users of the DiCoInfo, namely BA-students at Aarhus University, taking the course *Translating and editing texts for corporate websites*.[3] Seven students (out of the entire group of 10) participated in our test. To test the possible impact of the user-profile on the dictionary function (text production tasks in L2 or in L1 respectively), we also decided to gather feedback from 3 French Canadian translation students at the University of Montreal.

[3] The course is designed for Danish students of Business French in the second year of their study programme, and is aimed at enabling them to translate and edit Danish texts from corporate websites into French in the most appropriate way for the French market. As the students need multilingual terminology resources to complete their assignments, they are introduced to resources in the field, including the DiCoInfo.

Prior to the test, the participants were introduced to the functionalities of the DiCoInfo, including a presentation of its theoretical background and resources. The use of the DiCoInfo was illustrated by means of search sessions. The test was designed as a controlled questionnaire with specific instructions for the informants on how to perform the search task and complete the survey after each task (search results and search time), and fell into two distinctive parts: 1. Recording of information retrieval, and 2. Usability assessment. The first part consisted in performing four distinct search sessions mainly aimed at retrieving information from the new presentation in semantic classes. In the case of search session 1 – around the entry *fenêtre* – users were asked to perform search tasks and answer specific questions. The questions were designed to elicit data concerning the accessibility of the forms and the contents of the different information categories addressing *fenêtre*, particularly the new presentation of collocations and their meaning. The participants were asked to record the time spent on retrieving the information, and to assess the usability (access and user-friendliness) of the DiCoInfo. The second part is the evaluation part. It was designed to generate and elicit both quantitative and qualitative data (questions and answers, comments, explanations and suggestions concerning the information found in the DiCoInfo).

5.2 Results: Usability assessment and access and retrieval time scores

Immediately after each of the four search sessions the informants completed a survey in which they evaluated both the information they had found in the DiCoInfo and the ease of access to this information. The evaluation consisted in providing answers to several questions, of which the following two are of immediate concern to us: Was the information easy to understand? Was the information easy to find?

The numbers in each row in Table 4 below show the distribution of evaluation for each of the four search sessions on a 1-4 scale, 1 representing the highest degree of satisfaction and 4 the lowest. As each of the 10 informants answer two questions concerning every session, the total number of answers for each session is 20.

	SATISF. CAT 1	SATISF. CAT 2	SATISF. CAT 3	SATISF. CAT 4
SEARCH SESSION 1	3	14	3	0
SEARCH SESSION 2	0	19	1	0
SEARCH SESSION 3	0	11	9	0
SEARCH SESSION 4	9	11	0	0
% (rounded)	16%	69%	15%	0%

Table 4: Usability assessment

The majority of participants (69%) evaluate the usability of the DiCoInfo as satisfactory; none of them express strong dissatisfaction, while two minority groups express strong satisfaction (16%) or moderate dissatisfaction (15%). The results point to a high degree of overall perceived satisfaction (83%), but also indicate that there is room for improvement of overall usability. Table 5 below shows the highest, lowest, and average time scores for each of the four search sequences:

	HIGHEST TIME SCORE	LOWEST TIME SCORE	AVERAGE TIME SCORE
SEARCH SESSION 1	7 min.	0.5 min.	2.8 min.
SEARCH SESSION 2	7 min.	0.5 min.	2.5 min.
SEARCH SESSION 3	5 min.	1 min.	2.3 min.
SEARCH SESSION 4	4 min.	0.5 min.	2.8 min.

Table 5: Access and retrieval time scores

The time scores are flatly distributed across the search sequences in each of the three categories. There is, however, a striking gap between the highest and the lowest scores, as a few users (the ones expressing low satisfaction) spent about 14 times as much time as the fastest users.

5.3 Qualitative comments

The results of the analysis of the qualitative comments can be summarized as follows:

- Positive: All users emphasize the wealth and high quality of information, and acknowledge the usability of the DiCoInfo as a professional reference work for text production assistance.
- Negative: Almost all users stress the fact that they find it difficult to navigate within the articles, and express their experience in typical statements like "*c'est un peu difficile de s'y retrouver*" (= it's a bit difficult to find one's way), "*on s'y perd facilement*" (= you can easily get lost). Users obviously take the wrong path and need to backtrack. A case in point is the presence of two or more distinctive term records (*interface*), which adds to the difficulty of navigation.
- Suggestions for improvement: The use of distinctive colors to highlight sections, a clearer graphical layout, faster reloading of pages, and an even more efficient search-engine are some of the suggested features.

5.4 A paradoxical situation

Despite being a modest pilot-study, the test and feedback from users indicate that users with a semi-expert profile do appreciate the new resources. Access and retrieval is successful for most of them, provided they have been instructed in the use of the resource. There are, however, a few reservations to the positive output: some users fail to retrieve the information or spend much more time doing it; most users express frustration with the insufficient user-friendliness of navigation and data presentation. In fact, time scores in absolute figures (average time being almost 2.5 minutes!) reveal that speed of search and retrieval is influenced by presentation constraints and problems in understanding the rationale behind it. In short, the test reveals a paradoxical situation. On the one hand, users acknowledge the value of the information and perceive the DiCoInfo as a powerful instrument. On the other hand, using the DiCoInfo appears to be time consuming and demanding. This brings up the question of the lexicographical return on investment. We believe this return to be high, but wish it was even higher. There is still work to be done to achieve better output from the new features. One way to do this could be the integration of better, interactive instructions.

6 Conclusion: Status, future work and challenges

Up till now, approximately 300 LFs have been classified into wide-ranging semantic classes (generic and intermediate). Most represent generalizations over recurrent relationships in the DiCoInfo. Still, more LFs (most of them being non-standard LFs) need to be sorted out and organized within the hierarchy. Future work encompasses the selection of appropriate names for classes (i.e. names that clearly indicate what is comprised under it). On the software side, future work includes testing inheritance and triggering mechanisms in the hierarchy, and adapting the search engine to the new search options. In order to improve usability, we plan to add interactive instructions as well as a graphical search interface. Also, in the future it would be interesting to design and conduct usability tests of the DiCoInfo in real text production situations, and include log records of the number of clicks. This would generate an even more accurate picture of the overall usability of the DiCoInfo and help us in the ongoing development of user adaptive data presentation and access solutions. Finally, the work will have to be adapted to English and Spanish versions of the dictionary.

References

Dictionaries

Benson, M., I. Benson, I. & R. Ilson. 1997. *The BBI dictionary of English word combinations*. Amsterdam/Philadelphia: John Benjamins.

Binon, J., S. Verlinde, J. Van Dyck & A. Bertels 2009. *Dictionnaire d'apprentissage du français des affaires. Dictionnaire de compréhension et de production de la langue des affaires*. http://www.kuleuven.be/ilt/blf [last accessed: 16 June 2009]

DiCE = Alonso Ramos, M. 2011. *Diccionario de Colocaciones del Español*. Universidade da Coruña. http://www.dicesp.com [last accessed: 29 June 2011]

Hill, J. & Lewis, M. 1997, 2002. *LTP Dictionary of selected collocations*. Boston: LTP.

L'Homme, M. C. 2011. *Le DiCoInfo. Dictionnaire fondamental de l'informatique et de l'Internet*. http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi [last accessed: 15 March 2011]

Le Fur, D. (dir.) 2007. *Dictionnaire des combinaisons de mots*. Paris: Le Robert

McIntosh, C. (dir). 2009. Oxford Collocations Dictionary for Students of English. 2nd Edition. Oxford: Oxford University Press.

Mel'čuk, I. et al. 1984-1999. *Dictionnaire explicatif et combinatoire du français contemporain*. Montréal: Presses de l'Université de Montréal.

Other references

Alonso Ramos, M. 2004. Elaboración del Diccionario de colocaciones en español y sus aplicaciones. In Bataner, P. & J. de Cesaris (eds.). *De Lexicographia. Actas del I Simposio internacional de Lexicografia*, 149–162, Barcelona, IULA-Edicions Petició.

Charest, S., É. Brunelle, J. Fontaine & B. Pelletier. 2007. Élaboration automatique d'un dictionnaire de cooccurrences grand public. In *Actes de TALN 2007*, 283–292.

Cinkova, S. & P. Hanks Wyndham. 2010. *Validation of Corpus Pattern Analysis – Assigning pattern numbers to random verb samples. Annotation Guidelines*. http://nlp.fi.muni.cz/project/cpa/CPA_valiman.pdf

Jousse, A.-L. 2007. Organisation des fonctions lexicales pour un meilleur accès à l'information dans le DiCo. In Loiseau, M. et al. (eds.). *Autour des langues et du langage: perspective pluridisciplinaire. Papiers sélectionnés du Colloque International des Étudiants Chercheurs en Didactique des Langues et en Linguistique*, P.U.G.

Jousse A.-L. 2010. *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. PhD dissertation. Département de linguistique et de traduction, Université de Montréal & UFR de linguistique, Université Paris Diderot (Paris 7).

Jousse A.-L., A. Polguère & O. Tremblay. 2008. Du dictionnaire au site lexical pour l'enseignement/apprentissage du vocabulaire. In Grossmann F. and S. Plane (eds.). *Lexique et production verbale. Vers une meilleure intégration des apprentissages lexicaux*, Coll. Éducation et didactiques, 141-157. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Klosa, A. U. Schnörch & P. Storjohann. 2006. Elexiko — a lexical and lexicological, corpus-based hypertext information system at the Institut für deutsche Sprachemannheim. In *Proceedings of the 12th EURALEX International Congress*. 425-430, Turin, 6-9 September.

L'Homme, M.C. & P. Leroyer. 2009. Combining the semantics of collocations with situation-driven search paths in specialized dictionaries, *Terminology* 15(2), 258-283.

L'Homme, M.C., P. Leroyer & B. Robichaud. 2010. Advanced Encoding for Multilingual Access in a Terminological Database – A Matter of Balance, In *Terminology and Knowledge Engineering Conference 2010. Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges*, 12-13 August, Dublin.

Mel'čuk, I, A. Clas & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve (Belgique): Duculot / Aupelf - UREF.

Polguère, A. 2000. Towards a Theoretically-motivated General Public Dictionary of Semantic Derivations and Collocations for French. In *Proceedings of the 9th Euralex International Congress,* 517–527, Stuttgart, 8-12 August.

Polguère, A. 2003. Collocations et fonctions lexicales : pour un modèle d'apprentissage. In Grossmann F., and A. Tutin (eds.). *Les Collocations. Analyse et traitement*, 117-133, Amsterdam: De Werelt.

Tutin A. 2010. Les collocations dans les dictionnaires monolingues spécialisés de collocations. *Actes du 2e Congrès Mondial de Linguistique Française* (CMLF-2010).

Verlinde, S., T, Selva & J. Binon. 2006. The Base Lexicale du Français (BLF): A Multifunctional Online Database for Learners of French. In *Proceedings of the 12th Euralex International Congress*, 471–483, Turin, 6-9 September.

Grammemes

François Lareau

CLT, Macquarie University
Sydney, Australia
francois.lareau@mq.edu.au

Abstract

Importing key concepts from explanatory combinatorial lexicology, we revisit the notion of grammeme and show that it is an entity of the same level of abstraction as a vocable. Grammemes are polysemic and one of their acceptations is the basic one from which the others are derived. We propose criteria to identify it, and show how it can be used to guide the grouping of grammemes into inflectional categories.

Keywords

Grammemes, Grammatical Signs, Inflection, Morphology, Syntax.

1 Introduction

The inflectional system of a language is a hairy problem to tackle. To avoid getting lost, it is necessary to have solid theoretical ground to stand on. In this paper, we propose to further refine the definition of grammemes and inflectional categories within the Meaning-Text Theory (MTT) framework.[1] We will distinguish two things: grammemes proper, and grammatical units, i.e., their acceptations, and we will show how to identify the basic grammatical unit of a grammeme. We will also make a distinction between deep and superficial grammemes. Finally, we will propose a methodology for the study of inflectional systems that is based on explanatory and combinatorial lexicography. The analogy between the lexical and grammatical fields of study is twofold. First, the division of grammemes into grammatical units is similar to the division of vocables into lexical units. Second, the grouping of grammemes into inflectional categories, on the basis of their syntactics and semantics, is somehow reminescent of the grouping of vocables into parts of speech or lexical fields. But most importantly, it is the acknowledgment of the fact that polysemy plays just as central and confusing a role in the grammar as it does in the lexicon that underpins our approach.

[1] It is not the goal of this paper to look at new linguistic phenomena. For an illustration of how the concepts introduced here can be used in practice, see (Lareau, 2008, 2009).

2 The notions of grammeme and grammatical unit

2.1 Grammemes in the Meaning-Text Theory

(Mel'čuk, 1993) defines the grammeme as a *signification* that is part of an inflectional category, the latter being defined in the same book as the maximal set of significations that are mutually exclusive in a given (logical or semantic) position. The problem with this definition is that the concept of *signification*, although central in his theory of morphology, is defined nowhere in this book.[2] Is it a morphlogical or a semantic entity? Or is it a correspondence between such entities?

This purposefully vague definition aimed to resolve what we call the *grammeme polysemy paradox*. Indeed, grammemes tend to express several meanings. For instance, the future tense of Spanish verbs can either express a temporal meaning (*Vendré sobre las 6* 'I'll come around 6'), or a modal one (answering the question *¿Qué hora es?* 'what time is it?': *Serán las 2* 'it must be 2'—literally 'it will be 2'). In the former case, the future tense simply situates an event in time. In the latter however, the meaning is not at all temporal; the speaker is making a hypothesis about the *current* situation. There are clearly two distinct signs that we call *future tense* here. They have little in common semantically speaking, and as such, one would like to put them into different categories: the former with tenses, the latter with moods (or perhaps evidentiality). One would like to say that there are two grammemes **future** in Spanish. Yet, at the syntactic, morphological and phonological levels, these two signs are absolutely indistinguishable. So at the same time, one would want to say that there is only one grammeme **future** in Spanish.

This is probably what Mel'čuk had in mind when he wrote that a grammeme is a "beam of correspondences between a set of forms and a set of meanings" (Mel'čuk, 1993, p. 278). In other words, a grammeme is an interface between meanings and forms. He gives as example the plural of nouns in English, which, save for exceptions, is expressed by /s/, /z/ or /əz/. It can bear a number of meanings, for instance 'more than one X' (*The cats were sleeping there*), or 'all Xs' (*Piranhas are dangerous*).[3] One could of course describe this situation with a correspondence between every meaning and every form. But, if there are n meanings for a grammeme and it can be expressed by m forms, then one needs $n \times m$ correspondences. A more elegant and more cognitively plausible model is where the n meanings and m forms correspond to one entity of an intermediate level of representation, thus necessitating only $n + m$ correspondences.

Hence, Mel'čuk conceives the grammeme as an object of the syntactic representations that serves as an interface between semantics and surface syntax or morphology. As such, it is what we would like to call a "sub-semiotic" entity, i.e., the grammeme is not a linguistic sign, but rather something that sits on the correspondence path between meanings and forms. This resolves the apparent problem of the polysemy of grammemes, as it avoids saying that to each grammatical meaning corresponds one distinct grammeme on the one hand, and that grammemes are polysemic signs on the other hand, which would contradict the accepted definition of a linguistic sign in the MTT framework, where a sign can only have one meaning (again, this is discussed in (Mel'čuk, 1993)).

[2]The concept has later been discussed by (Polguère, 2008) and Mel'čuk (to appear).

[3]These semantic descriptions are of course very approximate; the definite/indefinite interfere here. It is not our topic to study the semantics of these grammemes; see for instance (Beyssade & Dobrovie-Sorin, 2005).

2.2 Grammatical units

Because of their polysemy, grammemes cannot be signs. This is similar to vocables in the lexical domain. Vocables are polysemic too, and they are not signs for the same reason: they are sets of lexical signs that share certain characteristics. The various acceptations of a polysemic vocable can have quite different meanings; for instance, the vocable **FACE** can denote a part of someone's head, or one of the surfaces of an object. These two meanings belong to two different lexical units (let us call them FACE$_1$ and FACE$_2$) and are different enough to be in different semantic fields: body parts for the former, and geometry for the latter. Yet, both FACE$_1$ and FACE$_2$ belong to the same vocable **FACE**, because they share non-trivial characteristics at every level of representation (even at the semantic level, their meanings are not entirely unrelated).

What we observe is that grammatical signs show the same kind of organization. Signs that share non-trivial characteristics are grouped under a unit of a higher level of abstraction, equivalent to the level of the vocable, and this is what we call *grammeme*. This abstract entity is a set of similar grammatical signs. To refer to the various acceptations of a given grammeme, we use the term *grammatical unit*, to echo the term *lexical unit*. This idea comes from (Kahane, 2002),[4] who viewed the grammatical units as "deep signs" just like lexical units, i.e., signs whose signified is a piece of the semantic representation, and whose signifier is a piece of the syntactic structure.[5] What distinguishes grammatical units from lexical units in (Kahane, 2002) is the nature of their signifier: grammatical units have as their signifier a grammeme, while lexical units have a vocable as their signifier.[6]

Hence, if we go back to the examples given above, there are in Spanish the grammatical units future$_1$ and future$_2$, which are two acceptations of the same grammeme **future**, in the same way as the lexical units FACE$_1$ and FACE$_2$ are acceptations of the vocable **FACE**.

2.3 Deep *vs* superficial grammemes

In MTT models, grammemes appear at the deep and surface syntactic levels. However, not all of them can appear at both levels. For example, definiteness in French is expressed by determiners. The signifier being a lexeme, it must have its own node at the surface syntactic level. Hence, in French, there are **definite** and **indefinite** grammemes only at the deep syntactic level; they are not needed anymore in surface syntax because their signifier has already been chosen. Therefore, we find it useful to distinguish between two types of grammemes:

Deep grammemes appear at the deep syntactic level and work as an interface between elements of the semantic representation (semantemes or communicative configurations) and elements of the surface syntactic level (function words or surface grammemes).

Superficial grammemes appear at the surface syntactic level and work as an interface between elements of the deep syntactic representation (deep grammemes or syntactic configurations) and elements of the deep morphological level (morphemes, prosodemes or word order).

It is around the deep grammemes that the grammatical system of a language is built, and for the rest of this paper, we refer to them simply as *grammemes*.

[4] He uses the word *grammie* in French, by analogy with *lexie*.

[5] (Kahane, 2002) makes a point of not having a deep syntactic level of representation.

[6] More precisely, for (Kahane, 2002), the signifier of a lexical unit is what he calls a *lexeme*, which corresponds more or less to what is usually referred to as *vocable* in the MTT litterature.

2.4 Inflection *vs* derivation: distinctive properties

Now, let us go back to Mel'čuk's definition of the grammeme. In (Mel'čuk, 1993), he defines it as an element of an inflectional category. Indeed, it is an important characteristic of grammemes that they are organized in inflectional categories: a grammeme is a grammeme only if it is opposed to other grammemes with which it forms a set of mutually exclusive elements. What makes these sets inflectional categories is, obviously, their inflectional nature. So let us review briefly the properties that distinguish inflection from derivation.[7]

Of all the properties mentionned by linguists to characterize inflection, as opposed to derivation, its obligatory nature is the one on which there is the largest consensus. (Jakobson, 1959), commenting on (Boas, 1938), said that the true difference between languages lies not in what they can express, but in what they force the speaker to express. This captures the essence of inflection. From this property follows the fact that inflectional meanings tend to be less numerous and more abstract. (Mel'čuk, 1993) notes six other characteristics that distinguish inflectional morphemes from the derivational ones: 1) they resist better to phraseologization, 2) they tend to have less restrictive combinatorics, 3) they tend to be expressed in a more regular way, 4) only grammemes can appear in agreement or concordance rules, 5) they tend to appear farther away from their lexical root, and 6) they do not modify the part of speech of the stem they attach to.

2.5 A methodology in two steps for the study of grammatical signs

The above characteristics help in identifying the grammemes of a language. But at the same time, these grammemes must be grouped into inflectional categories. As we have seen earlier, a grammeme may very well correspond to several meanings. In fact, one of the main problems in the study of grammemes is precisely their polysemy, often rich and subtle. It is hard to build a coherent model of a language's grammatical system if one tries to describe at once all the acceptations of a given grammeme. To make an analogy with lexicography again, this would amount to describing the meaning of a vocable without distinguishing its various acceptations.

Another pitfall in the study of grammatical signs is the unsuspected phenomenon of phraseology. Just like phrases can be lexicalized, combinations of grammemes can take on non-compositional meanings. To our knowledge, within the MTT framework, (Beck, 2007) was the first to mention this phenomenon, with examples of morphological phrasemes from Totonac. A more detailed account was later published as (Beck & Mel'čuk, 2011). We also gave examples of such phrasemes as well as grammatical collocations in French in (Lareau, 2008, 2009). Phraseologized expressions must be left aside when identifying the grammemes and the categories they belong to, in the same way that, for instance, the phraseme ⌜BY AND LARGE⌝ is irrelevant to the description of the lexeme LARGE.

In the following sections, we propose a resolutely discrete approach, in the sense that we believe it possible to isolate the acceptations of a grammeme, for the description of grammatical signs. In section 3, we discuss the notion of *basic grammatical unit* and propose a methodology to identify it. Then, in section 4, we propose principles for the grouping of grammemes into inflectional categories, based only on their basic grammatical unit.

[7] We use the term *inflection* in a broad sense that includes not only morphological inflection but also analytical forms such as auxiliaries and other function words.

3 The basic grammatical unit of a grammeme

Looking at the meanings a grammeme can express, we perceive intuitively that one is more sallient than the others. Grammemes have, like vocables, a basic sense from which its other meanings are derived somehow. This idea is not new, it was already expressed, for instance, in (Bello, 1847). We will call *basic grammatical unit* the acceptation of a grammeme that corresponds to its "proper meaning", by analogy with the *basic lexical unit* of vocables (Mel'čuk et al., 1995). It is often intuitively obvious what is the basic acceptation of a grammeme. Yet, it is difficult to formulate perfectly clear and rigorous criteria that systematically identify the basic grammatical unit of all grammemes. Below is our best attempt at it.

There are logically two types of criteria one can imagine: those based on the meaning of grammatical units, and those based on their combinatorics. It is not possible to have criteria based on the third component of linguistic signs, form, because, by definition, all acceptations of a grammeme are associated with the same forms.

3.1 Semantic criteria

If asked out of the blue what *he will eat* means, one would expect a native speaker to mention that the activity happens in the future. Then, if he thinks about it for a while, if the verb is put in various contexts, he might find other meanings to the future tense, but these meanings do not come to mind easily. So, trivially, the first criterion we could imagine, and in fact the one that best captures the essence of what we want to call *basic grammatical unit*, is the following:

Spontaneous interpretation criterion

The most common interpretation of a grammeme, the one that spontaneously comes to mind out of context, is the one that identifies the basic grammatical unit of a grammeme.

Obviously, this criteria opens the door to a certain subjectivity and must be used carefully. Besides, it does not always apply. It cannot easily be used for grammemes that do not correspond to semantemes (for example, grammemes of agreement). Even when a grammeme has acceptations that can be modeled with semantemes, the situation can be blurred by the fact that some grammatical meanings are "marked" while others are not. For example, in the tense system of English, the future and the past are marked, but not the present. A marked meaning being more sallient, it is more accessible to the speaker, while unmarked ones can easily go unnoticed. Hence, if a speaker is asked out of context what *he eats* means, she would probably define 'eat' rather than explain what the present tense means.

Empirically, we observe that, as is the case for the basic lexical unit of a vocable, the basic sense of a grammeme is often included in the other meanings of the same grammeme. In particular, there can be metaphorical relations between the basic meaning and the derived meanings of a grammeme. For example, the semantic relation between the present progressive that denotes a process taking place now (*I'm eating*—$\textsf{progressive}_1$) and the one that denotes a programmed action (*I'm leaving tomorrow*—$\textsf{progressive}_2$) could be described as a metaphor: '$X \oplus \textsf{progressive}_2$' $\approx$ 'X is so inexorably programmed that it is as if $X \oplus \textsf{progressive}_1$'. Based on this observation, we can formulate a second criterion:

Semantic inclusion criterion

If grammatical units A and B correspond to the same grammeme G and the meaning of A is included in that of B (by simple inclusion—direct or indirect—or via a metaphorical relation), then A is the basic grammatical unit of G.

(Mel'čuk et al., 1995) use the same criterion in the lexical domain. It often suffices, but there are cases where it does not identify what we would intuitively like to call the basic meaning of a grammeme. There are logically two cases where it would not work: 1) the criterion is not applicable because the meanings of the grammatical units under consideration are not included in one another, or 2) it is the basic sense of the grammeme that includes another. We have not found any example of the latter case, which would be a counter-example to our criterion, but it is easy to find examples of the former. For example, the *irrealis* sense of the English past tense:

- '$X \oplus \text{past}_1$' $\approx$ 'X happens before now' (*I had money last year*).
- '$X \oplus \text{past}_2$' $\approx$ 'condition X does not hold true [and I know it]' (*If I had money, I'd travel*).

Intuitively, we perceive the first one as the basic acceptation of **past**, but our criterion cannot be used here because there is no obvious inclusion relation between the two senses.

Generally speaking, criteria based on meaning pose two problems. First, not all grammemes are associated with meanings; notably all the "syntactic grammemes" of (Mel'čuk, 1994). Obviously, semantics-based criteria are useless for these. Second, while some grammatical meanings can relatively easily be defined by a semantic decomposition, as we normally do for lexemes, it is far from obvious that all meaning-bearing grammemes can. For example, the Japanese suffix –WA, which marks the theme of a sentence, has a signified that can only be described in the communicative structure, not by semantemes. For such grammemes, the semantic criteria fail.

Let us now turn to the criteria based on combinatorial properties.

3.2 Morphosyntactic criteria

In general, one would expect the basic acceptation of a grammeme to have less restrictive combinatorics than that of other acceptations. Given our conception of the deep grammeme as an interface device between semantics and surface syntax, the combinatorics of grammatical units cannot differ beyond the deep syntactic level. The combinatorial properties to consider are thus limited to two types: the syntactic configurations in which a grammeme can appear when it expresses a given meaning, and the lexical units with which it can combine.

First, it seems sensible to exploit the priviledged position of the root of a tree:

Syntactic root criterion (provisional)

If the grammeme under consideration combines with a class of lexemes that can be the syntactic root of a sentence, then its basic acceptation can be used in that position.

However, there are two problems with this criterion. First, there are grammemes that do combine with lexemes which can be syntactic roots, but that cannot appear at all in that position. For example, the past participle in English, although it combines with verbs (which are normally the root of a sentence), cannot appear in such a position, in any of its acceptations. Second, there are cases where this criterion identifies the wrong basic acceptation. In French, the grammeme

subjunctive usually appears on a subordinated verb (*Je veux qu'il aille demande à son chef* 'I want that he go ask his chief'). It can only be used on the root of the sentence if it bears an imperative meaning: *Qu'il aille demander à son chef !*, literally 'That he go ask his chief!'. Yet, one would not want to say that this imperative acceptation is the basic one for **subjunctive**. To avoid these problems, we need a more general criterion:

Syntactic polyvalence criterion
The basic sense of a grammeme is the one that is used in the most varied syntactic contexts.

For example, the imperative acceptation of the French **subjunctive**, though it can be used on the root of the sentence, can only be used in that position. In contrast, when this grammeme expresses subordination, it can occupy various positions: subject of a verb, complement of a verb, complement of a conjunction, etc.

Finally, the same criteria applies to the ability of grammemes to combine with different lexemes:

Lexical polyvalence criterion
The basic sense of a grammeme is the one that can be used in the most varied lexical contexts.

This criteria works only if there is a semantic incompatibility between a sense of a grammeme and certain lexemes. If the incompatibility is syntactic or morphological, then all acceptations of the grammeme are affected, since they all behave in the same way beyond deep syntax.

Generally speaking, the criteria based on combinatorics have a limited use since all acceptations of a grammeme correspond to the same object of the deep syntactic representation. Thus, it is only in the interface between semantics and syntax that one can observe differences in the combinatorics of various acceptations of a grammeme.

Finally, we want to insist on the fact that these criteria cannot, in isolation, systematically identify the basic acceptation of a grammeme. Together, however, they can lead the linguist when confronted with a non-obvious case.

Identifying the basic grammatical unit of a grammeme is a very important step in our methodology because it is this acceptation only that will be considered when grouping the grammemes into inflectional categories. Let us now turn to this problem.

4 Grouping grammemes into inflectional categories

As we have mentionned earlier, grammemes are only grammemes if they are part of an inflectional category. Then, what are the criteria that should guide the grouping of grammemes into categories? There are logically three major types of criteria that could be considered, based on the three components of signs: meaning, form and combinatorics. We do not know of any linguist who proposed building inflectional categories based on the forms of its members. Given the arbitrary nature of the signifiers of linguistic signs and the fact that grammemes are often expressed cumulatively, this avenue looks like a dead end. However, the other two components of signs seem viable options, so let us explore them.

But first, let us emphasize that it is the *deep* grammemes, and not the grammatical units nor the superficial grammemes, that we want to group into inflectional categories. It is indeed at this level that grammatical systems are organized.[8]

[8]This echoes the idea of (Kahane, 2009) that the deep syntactic level is where signs are organized in a sentence.

4.1 Morphosyntactic criteria

The combinatorics of grammatical signs offers an interesting basis for rigorous and verifiable criteria because it is relatively easy to observe. (Martinet, 1979) proposed a methodology essentially based on it, which (Touratier, 1996) pushed a little further. Both get similar results when applying their method to French, which shows its reproducibility. It relies mainly on two criteria, given in (Martinet, 1979), that we reformulate below with our terminology. These two principles correspond to the two facets of *contrastive distribution*, a key concept in linguistics.

Mutual exclusion criterion

Grammemes of a same inflectional category are mutually exclusive.

Combinatorial similarity criterion

Grammemes of a same inflectional category have similar combinatorics.

Thus, for instance, the French adjectival grammemes **masculine** and **feminine** belong to the same category because they are mutually exclusive and they have identical combinatorial properties: both combine with adjectives and with grammemes of number.

However, we think that combinatorics-based criteria alone do not suffice. (Touratier, 1996) made a model of French verb conjugation based strictly on the combinatorics of grammatical signs, and his results illustrate very well the limitations of this method. His reasoning goes roughly as follows. First, the imperfect can combine with the subjunctive (which gives the subjunctive imperfect), while the past tense cannot. Like the subjunctive, the future tense can also combine with the imperfect (which gives the conditional)[9] but not with the past tense. Therefore, the future is part of the same category as the subjunctive since they have the same combinatorics, and the past, because they are mutually exclusive. Since the subjunctive cannot be anything else than a mood, the past and future must also be moods. We do not find this reasoning convincing. The fact that the future cannot combine with the past is indeed a hint that they belong to the same category, but their incompatibility with the subjunctive does not necessarily imply that they form a paradigm with it. It could simply be due to the fact that the two are tenses and that the subjunctive does not combine with tenses.[10]

We believe that criteria based on combinatorics alone do not suffice to group grammemes into categories, and that it is necessary to take into account the basic meaning of these grammemes (but only the basic one!).

4.2 Semantic criteria

(Mel'čuk, 1993) defines a category (a notion that englobes inflectional categories) as a set of elements that are mutually exclusive in a given "semantic or logical position". Thus, it is mainly (but not exclusively) on semantic criteria that he builds his inflectional categories. Following his definition of a category, one can formulate the following criterion:

[9] Indeed, the conditional in French (*mangerait*) is formed by the combination of two suffixes, –R (future tense) and –AIT (imperfect), in a way similar to the English conditional (*would eat*), which is expressed by the auxiliary WILL (future tense) in its past form.

[10] This would also mean that the imperfect is not a tense, an analysis argued for by (Vet, 2007) and (Lareau, 2008).

Semantic mutual exclusion criterion

The basic meanings of the grammemes of an inflectional category are mutually exclusive.

Here, it is important that we consider only the basic sense of grammemes, a point that (Mel'čuk, 1993) missed. Otherwise, we are forced to stipulate unnecessary elements of the syntactic and morphological representations that are simply traces of the polysemy of grammemes. For example, if we took into account all the senses of the conditional in English, we would have to distinguish at the syntactic and morphological levels at least two grammemes for the conditional: one that belongs to moods (for the conditional that expresses politeness, as in *Would you pass me the salt?*), and another that belongs to tenses (for the conditional that expresses temporal relations, as in *He told me he would come*). Yet, from the deep syntactic level, and up to the surface realization, both have exactly the same behaviour. The result is an unnecessary duplication of rules in every module of the grammar, a problem that we avoid if we take into account only the basic meaning of grammemes. Finally, we complement this criterion with one inspired from the definition of a vocable in (Mel'čuk et al., 1995):

Semantic similarity criterion

The basic acceptation of the grammemes of an inflectional category have an obvious semantic similarity.

This avoids putting in the same category the French past and subjunctive, like (Touratier, 1996).

5 Conclusion

We have distinguished between *deep grammemes*, which belong to the deep syntactic structure and are at the core of the grammatical system of a language, and *surface grammemes*, which are found in surface syntax. Grammemes, because they are polysemic, are not signs, but entities of the same level of abstraction as the vocables in the lexical domain. We call each acceptation of a grammeme a *grammatical unit*, and one of them constitutes the *basic grammatical unit* of that grammeme, a concept borrowed from lexicography. We have proposed some criteria based on the meaning and the combinatorics of grammatical units to identify the basic one for a grammeme. It is only the basic grammatical units that must be considered when grouping the grammemes into inflectional categories, and phraseologized grammatical signs must be ignored. Finally, we proposed criteria to decide whether two grammemes belong to the same category: their basic acceptations must be mutually exclusive and show certain similarities at every level of representation.

Acknowledgments

This paper is an extract of a Ph.D. dissertation we wrote under the supervision of Sylvain Kahane and Igor Mel'čuk, with whom we have discussed our ideas on countless occasions. Their feedback was gold. The whole thesis was also read and commented upon by Laurence Danlos, Pierre Le Goffic, Alain Polguère and Owen Rambow, whom we wish to thank again for their insightful comments. Finally, three anonymous reviewers have generously provided useful suggestions that led, we hope, to a better paper. This research was conducted mostly with the financial support of the SSHRC, through scholarship #752–02–1501.

Bibliography

Beck, D. 2007. Morphological phrasemes in totonacan inflection. In *Proceedings of MTT 2007*, 107–116. Klagenfurt.

Beck, D. & I. Mel'čuk. 2011. Morphological phrasemes and totonacan verbal morphology. *Linguistics*, 49(1):175–228.

Bello, A. 1982 [1847]. *Gramática de la lengua castellana*. Madrid: Edaf.

Beyssade, C. & C. Dobrovie-Sorin. 2005. *Définir les indéfinis*. Paris: CNRS Éditions.

Boas, F. 1938. Language. *General Anthropology*.

Jakobson, R. 1959. Boas' view of grammatical meaning. *American Anthropologist*, 61(5).

Kahane, S. 2002. *Grammaire d'Unification Sens-Texte : Vers un modèle mathématique articulé de la langue*. Document de synthèse pour l'habilitation à diriger les recherches, U. Paris 7.

Kahane, S. 2009. Defining the deep syntactic structure: How the significant units combine. In *Proceedings of MTT 2009*, 199–211. Montréal.

Lareau, F. 2008. *Vers une grammaire d'unification Sens-Texte du français: le temps verbal dans l'interface sémantique-syntaxe*. Thèse de doctorat, U. de Montréal/U. Paris 7.

Lareau, F. 2009. Le temps verbal dans l'interface sémantique-syntaxe du français. In *Proceedings of MTT 2009*, 223–232. Montréal.

Martinet, A. 1979. *Grammaire fonctionnelle du français*. Paris: CREDIF.

Mel'čuk, I. 1993. *Cours de morphologie générale. Introduction et première partie: le mot*, volume 1. Montréal/Paris: Presses de l'Université de Montréal/CNRS Éditions.

Mel'čuk, I. 1994. *Cours de morphologie générale. Deuxième partie: significations morphologiques*, volume 2. Montréal/Paris: Presses de l'Université de Montréal/CNRS Éditions.

Mel'čuk, I. To appear. *Semantics: from Meaning to Text*, volume 1. Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, I., A. Clas & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Polguère, A. 2008. *Lexicologie et sémantique lexicale* (2nd edition). Montréal: Presses de l'Université de Montréal.

Touratier, C. 1996. *Le système verbal français : description morphologique et morphématique*. Paris: Armand Colin.

Vet, C. 2007. The descriptive inadequacy of reichenbach's tense system: a new proposal. In de Saussure, L., J. Moeschler & G. Puskas (eds). *Tense, Mood and Aspect: Theoretical and Descriptive Issues*, 7–26. Amsterdam/New York: Rodopi.

ILexicOn: toward an ECD-compliant interlingual lexical ontology described with semantic web formalisms

Maxime Lefrançois, Fabien Gandon

EPI Edelweiss – INRIA Sophia Antipolis
2004 rt des Lucioles, BP93, Sophia Antipolis, 06902, France
Maxime.Lefrancois@inria.fr | Fabien.Gandon@inria.fr

Abstract

We are interested in bridging the world of natural language and the world of the semantic web in particular to support natural multilingual access to the web of data. In this paper we introduce a new type of lexical ontology called *interlingual lexical ontology* (ILexicOn), which uses semantic web formalisms to make each interlingual lexical unit class (ILU^c) support the projection of its semantic decomposition on itself. After a short overview of existing lexical ontologies, we briefly introduce the semantic web formalisms we use. We then present the three layered architecture of our approach: i) *the interlingual lexical meta-ontology* (ILexiMOn); ii) the ILexicOn where ILU^cs are formally defined; iii) the data layer. We illustrate our approach with a standalone ILexicOn, and introduce and explain a concise human-readable notation to represent ILexicOns. Finally, we show how semantic web formalisms enable the projection of a semantic decomposition on the decomposed ILU^c.

Keywords

Explanatory Combinatorial Lexicology; Semantic Web; Semantics; Semantic decomposition; Conceptual layer of representation; Conceptual participant slots; Interlingual Lexical Primitives.

1 Introduction

In this paper we introduce and illustrate the core of the ongoing ULiS project that is at the barycenter of the Meaning-Text Theory (MTT), pivot-based NLP techniques, and the semantic web formalisms. What we aim for in the ULiS project is a *universal linguistic system* (ULiS), through which multiple actors could interact with *interlingual knowledge bases* in multiple controlled (i.e., restricted and formal) natural languages. Each controlled natural language (dictionary, grammar rules) would be described in a part of a *universal linguistic knowledge base* (ULK). Besides this, the ULK consists in one specific interlingual knowledge base. Actors could then enhance their controlled natural language through different

actions in controlled natural language (e.g., create, describe, modify, merge, or delete lexical units in the dictionaries and grammar rules; connect situational lexical units to interlingual lexical units; add linguistic attributes with their associated rules, etc.) These actions are assigned the top-priority as the universal linguistic knowledge base would be the cornerstone of the universal linguistic system.

The aim of this paper is to introduce the core of such a universal linguistic knowledge base, i.e., the *interlingual lexical ontology* (ILexicOn). Roughly, we aim to port pure semantic features of *explanatory combinatorial dictionaries* (ECD) to the semantic web formalisms.

The rest of this paper is organized as follows. Section 2 surveys the related work on lexical ontologies and interlingual lexical ontologies. Due to the novelty of our approach, we chose to develop a section on Semantic Web formalisms (Section 3), and to focus on one specific feature of our model: the formal definition of the *interlingual lexical unit classes* (ILU^cs, Section 4). We give an overview and illustration on the architecture of our model (subsection 4.1), then we justify our novel approach for the lexicographic definition of ILU^cs and introduce the modeling choices that we made and the notations that we use (Subsection 4.2). We will leave the study of lexical functions and the description of what is not interlingual for a next paper.

2 Related work

Lexical ontologies, i.e., an ontology of lexical(-ized) concepts, are widely used to model lexical semantics. There exist many of them. Some have broad coverage but shallow treatment (i.e., with no or little axiomatization) such as Princeton WordNet (e.g., Miller et al., 1990), Euro-WordNet (Vossen, 1998), and some have small coverage but are highly axiomatized such as CYC (Lenat et. al., 1990), SUMO (Lenat et al., 1998), DOLCE (Niles & Pease, 2001), Mikrokosmos (Nirenburg et al., 1996), HowNet / E-HowNet (Dong & Dong, 2006), FrameNet (Baker et al. 1998). They use different theories of lexical semantics, but only one of them is ECD-compliant: the Lexical System (Polguère, 2009) and it focuses only on the representation of lexical functions, and does not define lexical units nor uses semantic web formalisms.

On the other hand, the Universal Networking Language (UNL) is a meaning representation language, originally designed for pivot techniques Machine Translation. Its dictionary is an interlingual lexical ontology based on so-called Universal Words, but the lack of argument frames and lexical functions in the UNL dictionary was pointed out in (Bogulsavsky, 2002, Bogulsavsky, 2005). To the best of our knowledge, this is when the idea of an ECD-compliant interlingual lexical ontology was first mentioned. After the semantic web formalisms were introduced at the W3C, an attempt to port the UNL to semantic web formalisms was the topic of a W3C incubator group led by the inventor of UNL: H. Uchida (XGR-CWL, 2008), but no improvement was made to the lexical ontology.

Benefits of using semantic web formalisms are high as it enables us to construct an axiomatized graph-representation of a lexical ontology, with validation and inference rules. This is why we propose to use semantic web formalisms to model an ECD-compliant interlingual lexical ontology.

3 The Semantic Web formalisms

The semantic web stack consists in a set of World Wide Web Consortium (W3C) recommendations. These recommendations propose: i) a unified data structure (RDF Graphs); ii) corresponding query/update language and protocol (SPARQL); iii) fragments of logics with different expressivity to capture formal semantics of the data schemas (RDFS, OWL); and iv) a rule language offering an alternative for capturing inferences over the data (RIF). In this paper, we show how suitable this framework is to design an ECD-compliant ILexicOn.

Universal Resource Identifier (URI). Broadly, URIs may be assigned to anything we want to talk about. Universal Resource Locators (URLs) are specific URIs that identify and locate resources on the web. That said, URIs are meant not only to identify Web Documents, but any resource, including real-world objects, interlingual lexical unit classes (ILU^cs), interlingual lexical unit instances (ILU^is) and interlingual semantic relations (ISemRels). For instance, the URI of the ILU^c corresponding to the English LU KILL$^{1.1}$ (numbered according to the Longman Dictionary of Contemporary English) may be identified as: http://ns.inria.fr/ulk/2011/06/10/ilexicon-ex#Kill1.1, or ilexicon:Kill1.1 using a namespace prefix.

Resource Description Framework (RDF). RDF models directed labeled multigraphs that serve as a base structure for the semantic web stack of the W3C, together with the URIs. RDF enables the description and connection of resources which can be anonymous resources or resources identified by an URI. In RDF, the atomic piece of knowledge is the triple of the form (subject, predicate, object) with predicate being an rdf:Property. For instance, the assertion "John kills Mary" may be decomposed in three RDF triples: (ex:k01, rdf:type, ilexicon:Kill1.1), (ex:k01, ilexicon:hasAgent, ex:John01) and (ex:k01, ilexicon:hasKilled, ex:Mary01)

Sitting at the bottom of the recommendation stack, RDF imposes an open world assumption to the whole semantic web stack. In particular, the types of resources (Classes) and links (Properties) are only constrained by the fact they should be valid URIs. Note that open world assumption implies that one can reuse or extend anyone's knowledge base, and assert anything on anything.

Resource Description Framework Schema (RDFS). RDFS stands for RDF schema and allows us to declare hierarchies of classes to type the RDF graphs, in other words lightweight formal ontologies. A schema in RDFS enables us to associate a class to existing resources, a type to the relationship between existing instances of these classes. It also enables us to define domain (resp. range) of the relation, i.e., the class to which subjects (resp. objects) of the relation belong to. RDFS defines inferences to be applied using these hierarchies of types and the signatures of properties. By allowing us to provide URIs to types, RDFS enables the description of the taxonomic skeleton of a lightweight ontology in a universal language, with universal identifiers and semantics (with simple axioms e.g., subClassOf, subPropertyOf).

Ontology Web Language (OWL). OWL is a meta-language that roughly speaking extends RDFS to enable us to describe ontologies with additional logical expressivity. In an ontology, resources are divided in three sets: classes, individuals that populate these classes, and properties that link those individuals. Also, depending on whether we want less complexity or

more expressiveness, OWL recommends the use of more or fewer constructors for classes and properties (e.g., intersection, union, cardinality restriction, etc.).

SPARQL. SPARQL is the RDF query/update language and protocol.

4 ILexicOn: The Interlingual Lexical Ontology

Now that we have positioned our work and introduced the semantic web formalisms, we present the focus of this paper: the *Interlingual Lexical Ontology* (ILexicOn). Roughly, the ILexicOn contains the pure semantic features of the *Explanatory Combinatorial Dictionary* (ECD).

4.1 Overview

Our approach is based on a three layered architecture:

1. **The meta-ontology layer: the interlingual lexical meta-ontology (ILexiMOn)**. It is the schema that every ILexicOn must satisfy. We designed a light core-ILexiMOn[1] that is illustrated on Figure 1.

2. **The ontology layer: the interlingual lexical ontology (ILexicOn).** The ILexicOn contains the formal definitions of the *interlingual lexical unit classes*, called ILU^{c}s, which are instances of the ILexicalUnit meta-class from the core-ILexiMOn. The ILexicOn contains also the definition of the *interlingual semantic relations*, called ISemRel, that are instances of the ISemRelation meta-class from the core-ILexiMOn. To illustrate our approach, we designed a light standalone ILexicOn[2]. A few ILU^{c}s are illustrated on Figure 1, and the whole ILexicOn is illustrated on Figure 2. To concisely describe the whole ILexicOn on Figure 2, we adopted a notation inspired from Sowa's conceptual graphs (Sowa, 1984), and detailed in the section 4.3. Let us just say that each rectangle is the definition place of the ILU^c that is written in its top-left corner.

3. **The data layer: the interlingual semantic representations (ISemR).** The data layer contains *interlingual semantic representations* (ISemR). Nodes are *interlingual lexical unit instances* (ILU^{i}s), and arcs are *interlingual semantic relations* (ISemRels). This layer is illustrated in Figure 1, and we illustrated our approach with three simple ISemRs[3] on Figure 2.

Figure 1 illustrates the architecture of our work, with its integration in the semantic web formalisms. From top to bottom: 1) the semantic web formalisms, with a few OWL classes and properties that are useful for our work; 2) the detailed core-ILexiMOn; 3) an overview of the ILexicOn we detail in Figure 2; and 4) an overview of the data layer.

[1,2,3] RDF/XML documents are available at URLs:
http://ns.inria.fr/ulk/2011/06/10/ileximon-core. For the core-ILexiMOn
http://ns.inria.fr/ulk/2011/06/10/ilexicon-ex. For the light ILexicOn.
http://ns.inria.fr/ulk/2011/06/10/sems-ex. For the data layer.

Notice that: i) ILU^{i}s from the data-layer are instances of ILU^{c}s described in the ILexicOn, that are themselves instances of the ILexicalUnit meta-classes described in the ILexiMOn; and ii) properties used to link two resources in a layer are described in an upper layer.

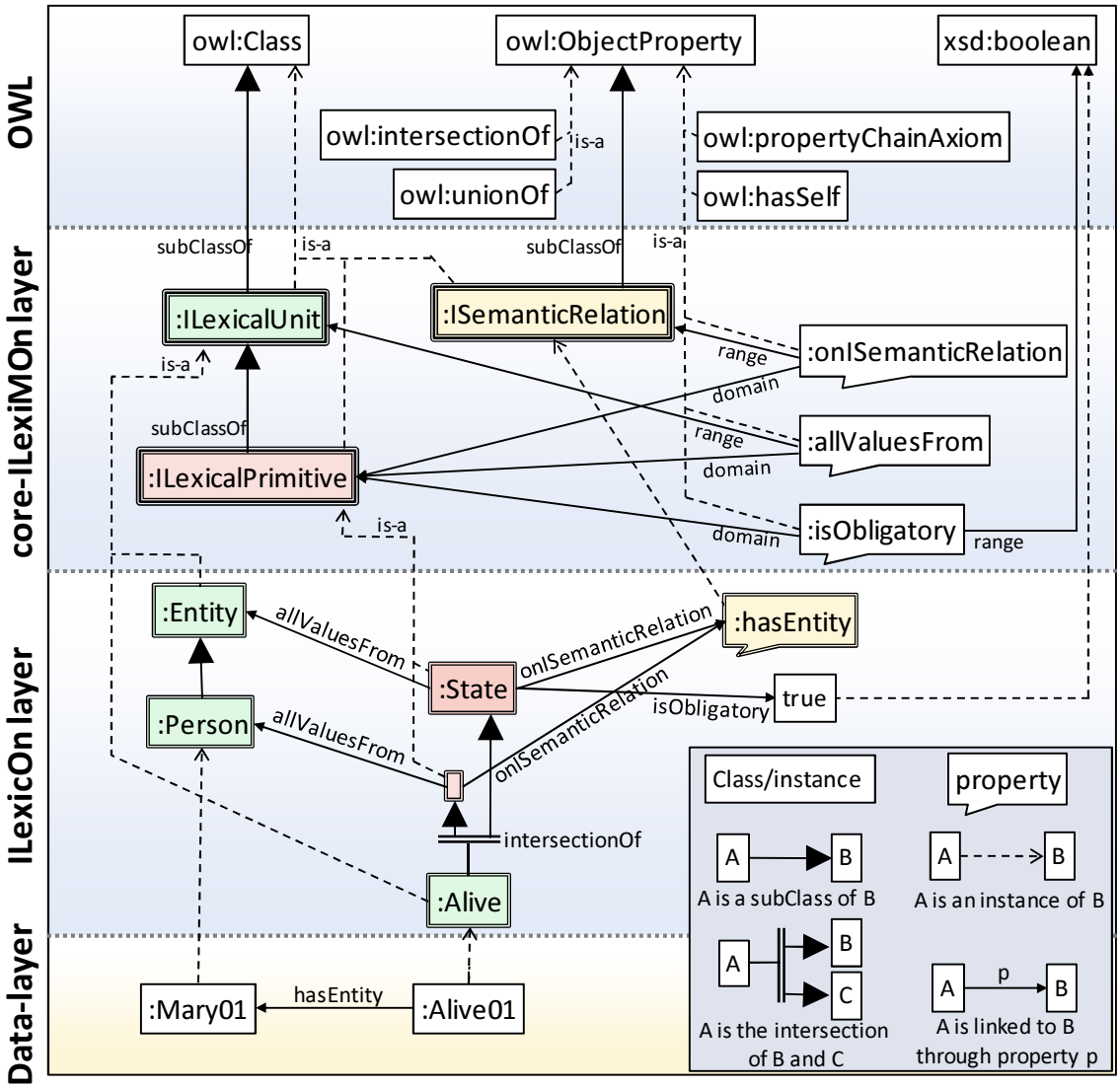


Figure 1: The three layered architecture of our work, with details of the core-ILexiMOn and overview of the ILexicOn and the data-layer.

Semantic web formalisms are truly well-suited for the design of an ECD-compliant lexical ontology. Indeed, the chosen architecture with a meta-level ensures to satisfy the three construction principles of an ECD out of the four specified in (Mel'čuk et al., 1995). Firstly an ILexicOn is bound to be explicit, to comply with the ILexiMOn and to be internally coherent (formality and internal coherence principles). Furthermore, all descendants of an ILU^{c} inherit some of its features, ensuring uniformity (uniformity processing principle). On the other hand, the sufficiency principle can't be fully ensured, but adding rules in the ILexiMOn may contribute to satisfy this principle by providing means to infer new information and/or to highlight missing information.

4.2 A novel approach for the lexicographic definition of lexical units

4.2.1 ILexicOn in the conceptual layer of representation

To notate differently ILU^cs and ILU^is avoids confusing ILUs appearing in the lexicon and ILUs in use in the semantic representation of an utterance. In the MTT, two kind of lexicographic definitions of a LU are thought: i) in some natural language (i.e., in the surface phonologic layer of representation), or ii) using a semantic representation format (i.e., in the semantic layer of representation). We claim that both approaches consist in generically instantiating (or constructing) a semantic decomposition of the ILU^c. In our approach, we clearly want to separate out the ILexicOn layer and the ISem layer. We therefore propose ways to represent the lexicographic definition of an ILU^c without ILU^i, nor the semantic representation of its semantic decomposition.

The main proposal of this article is thus to raise the lexicographic description of an ILU^c to the ILexicOn layer. As this layer is deeper than the semantic representation layer, we propose to consider it in the *conceptual layer of representation* and thus use the notion of *linguistic situation denoted by a* ILU^c *L*, i.e., SIT(L) as the union of semantic decompositions of L, and the notion of *participant of SIT(L)* for each node in SIT(L). A participant of SIT(L) may be obligatory or optional (Mel'čuk, 2004).

> ***Notations:*** Let L be an ILU^c, and $\mathbf{L}=\{L_i\}$ be the set of ILU^cs of the minimal semantic decomposition of L.

L is a subset of the set of *participants* of SIT(L). Also, one of the L_i is the ILU^c which summarizes the meaning of the decomposed ILU^c. The definition we gave to SIT(L) and participants of SIT(L) is compatible with the MTT *participant inheritance principle* that states (Mel'čuk, 2004):

> SIT(L) inherits all obligatory participants of all SIT(L_i) that correspond to the predicative meanings of $^{(}L_i^{)}$ (i.e., ILU^c_i) which compose the meaning $^{(}L^{)}$ (i.e., ILU^c).

We thus propose a novel approach to the lexicographic definition of an ILU^c that consists in projecting the minimal semantic decomposition of the ILU^c *on* the ILU^c using Semantic Actant-like slots.

4.2.2 Interlingual lexical units (classes and instances) and interlingual semantic relations

ILU^cs are instances of the ILexicalUnit meta-class from the ILexiMOn (c.f., Figure 1). They are defined in the ILexicOn (c.f., Figure 2, e.g., Entity, Person, State, Alive, Event, Cause). In our notation, symbol < represents the rdfs:subClassOf axiom that may be used to state inheritance between ILU^cs (e.g., Person<Entity, Alive<State, Cause<Event). For instance, The ILU^c Person is a sub-class of the ILU^c class Entity, and the ILU^c Entity is the parent of the ILU^c Person. Complex ILU^cs may be constructed through owl:intersectionOf and owl:unionOf. Finally, *interlingual lexical unit instances* (ILU^is) are instances of ILU^cs and are used in the ISem layer as nodes of the interlingual semantic representations. At this point, one may ask

what an ILU[c] that inherits from no other ILU[c] is. *A priori*, such an ILU[c] is semantically void, and should therefore not be considered as a lexical primitive of the ILexicOn.

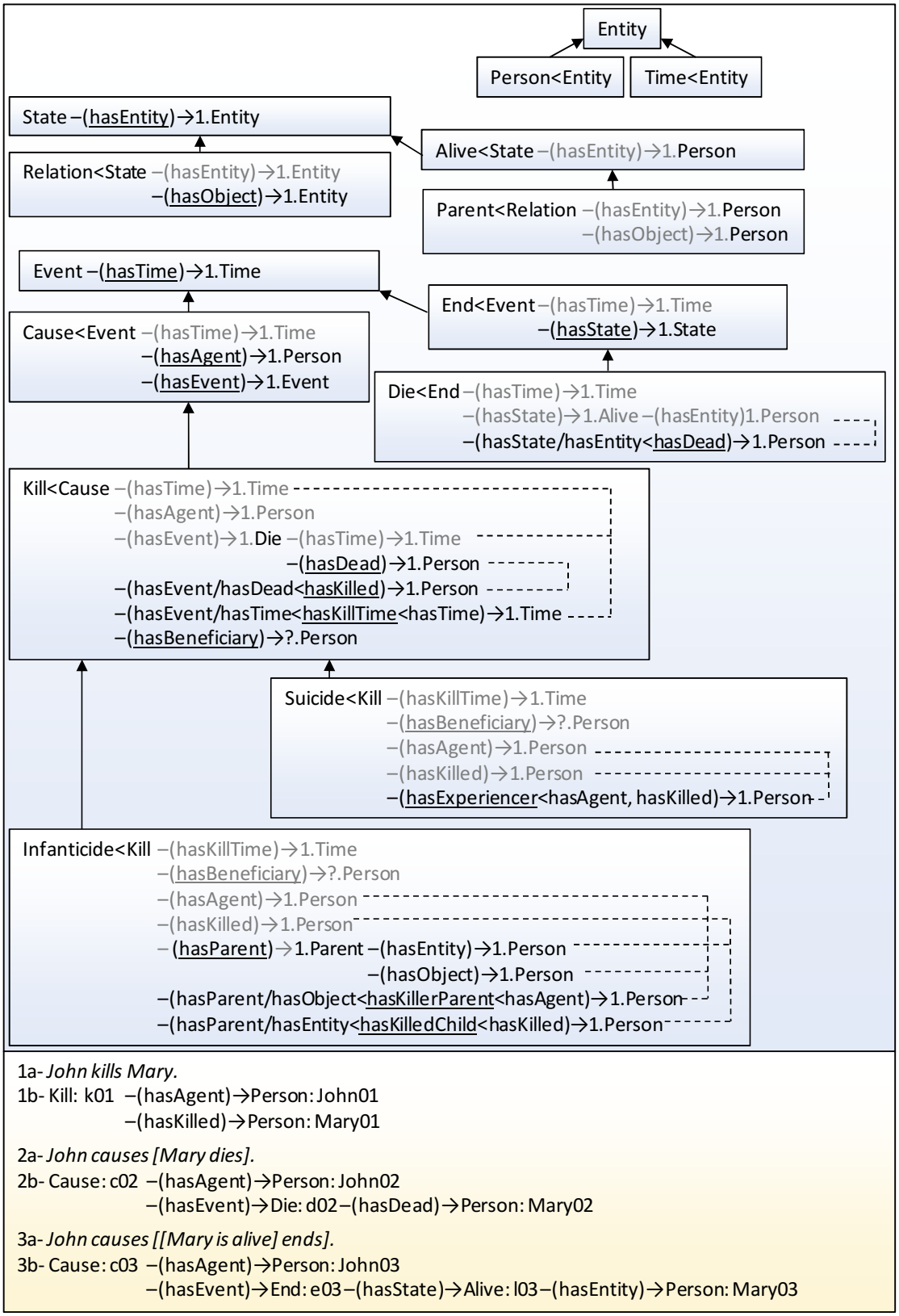


Figure 2: The light standalone ILexicOn and three ISemRs described with our new notation.

ISemRels are instances of the ISemRelation meta-class of the ILexiMOn, and thus instances of owl:ObjectProperties. They are introduced in the LexicOn and used in the data layer to link ILU^{i}s (see Figure 1&2). In our notation, symbol < represents the rdfs:subPropertyOf axiom that may be used to define a new ISemRel as being a sub-ISemRel of one or more ISemRels (e.g., hasExperiencer<hasAgent, hasKilled). Symbol / represents the owl:propertyChainAxiom axiom that may also be used to state that a ISemRel is a super-ISemRel of the composition of two or more ISemRels (e.g., hasState/hasEntity<hasDead). These two axioms may be combined to define complex ISemRels (e.g., hasEvent/hasTime<hasKillTime<hasTime).

4.2.3 From interlingual lexical primitives to projected minimal semantic decomposition.

As the ILexicOn that we designed is interlingual, we limit the scope of our study to purely semantic features of the ECD. Thus Semantic Actants are not considered as their definition relies on the definition of the expressibility of a participant in texts, which relies on non-semantic features (Mel'čuk, 2004). We introduce a new notion, i.e., *Conceptual Participant slots* (ConP-slot): the implicit link that exists between an ILU^c L and one of the participants of the minimal semantic decomposition of L.

We stated in Subsection 4.3.1 that an ILU^c that inherits from no other ILU^c is *a priori* semantically void, an ILU^c is semantically void. Yet we may precise our thought and introduce the *interlingual lexical primitive classes* (ILP^{c}s): an ILU^c L is a ILP^c if and only if it derives from no other ILU^c but has at least one ConP-slot. Non- lexical primitives then derive from one or more lexical primitives following the *ConP-slot* inheritance and introduction principle:

An ILU^c L inherits from its parents' ConP-slots, and may also introduce new ConP-slots;

This principle highly restricts the number of ConP-slots of L compared to the number of participants of L, indeed, one may consider only participants that are necessary and sufficient to the minimal projection of L. ILP^{c}s are defined as instances of the ILexicalPrimitive meta-class from the ILexiMOn (c.f., Figure 1). An ILP^c must be linked through: i) the onISemanticRelation property to exactly one ISemanticRelation; ii) the allValuesFrom property to exactly one ILexicalUnit; and iii) the isObligatory property to exactly one xsd:boolean.

In Figure 2, each line with an arrow in the definition of an ILU^c represents a conceptual participant slot (ConP-slot) that restricts the use of a specific ISemRel for this ILU^c and its descendants. Actually, such a line means that the defined ILU^c is a sub-class of an ILP^c. For instance, the line State–(hasEntity)→1.Entity states that any instance of the State class is linked exactly once through the hasEntity relation to an instance of the Entity class. Let us focus on the notation used on Figure 2:

- **Inheritance**. ConP-slots may be newly defined (black font, e.g., State–(hasEntity)→1.Entity), fully inherited (grey font, e.g., Relation<State–(hasEntity)→1.Entity) or partially inherited (grey font for the inherited part, e.g., Alive<State–(hasEntity)→1.Person). The ILU^c on the right hand side of the line is called the *current range of the ConP-slot*.

- **Obligatory vs. optional.** A ConP-slot may be obligatory (symbol 1, e.g., Alive<State–(hasEntity)→1.Person) or optional (symbol ?, e.g., Kill<Cause–(hasBeneficiary)→?.Person). When an optional ConP-slot is inherited, it may be restricted to being obligatory.

- **Domain/range of the ISemRel.** As an ISemRel is an rdf:Property, it may restrict its domain and its range i.e., what ILU^c the subject (resp. the object) of a triple that involves this ISemRel does belong to. When an ISemRel is underlined, it means that its domain is set to the defined ILU^c, and that its range is set to the current ILU^c range of the ConP-slot. (e.g., State–(hasEntity)→1.Entity).

- **ISemRel subproperty and composition axioms.** As we stated in section 4.2.2, complex ISemRel may be defined thanks to inheritance and composition. There are benefits in using such ISemRel to qualify a new ConP-slot. In fact, this combined with the maximum cardinality of ConP-slots restricted to 1, imposes the equality of ILU^i in the data-layer. We illustrate these inferable equalities by dotted lines on the right of ConP-slots.

The ISemRel inheritance and composition is what enables the projection not only of trees, but also graphs, onto one node. Thus, each ILU^c described in the ILexicOn contains the projection of its semantic decomposition graph. We illustrated this on Figure 2 with complex ILU^c such as ilexicon:Suicide (the killer is the killed person) and ilexicon:Infanticide (the killer is the parent of the killed person).

5 Conclusions and discussions

We introduced and illustrated a three layer architecture that describes ECD-compliant interlingual lexical ontologies using semantic web formalisms. We introduced the core of an interlingual lexical meta-ontology (ILexiMOn) that composes the top-layer of the architecture. This ILexiMOn describes the middle-layer interlingual lexical ontology called ILexicOn, where classes of interlingual lexical units (ILU^{c}s) are described. Finally interlingual semantic representations are part of the third layer. We introduced a novel approach to formally define ILU^{c}s: we make ILU^{c}s support a projection of their semantic decomposition, thus keeping their definition in the same conceptual layer of representation. We introduced a human-readable notation to represent ILexicOn, and we used this notation to illustrate our approach with a simple standalone ILexicOn. We thus showed how simple and complex ILU^{c}s may be formally defined with our novel approach.

On the basis of what is introduced in this paper, our research currently progresses in three directions: 1) how to model pure-semantic lexical functions in the ILexiMOn or in the ILexicOn (notice that the ILU^c ilexicon:End *is* a specific lexical function); 2) The formalization of validation and inference rules to validate and augment i) the ILexicOn, ii) an interlingual semantic representation (these rules will be included in the LexiMOn); 3) how to model what we call the situational lexical ontology that describes situational lexical units with their semantic actants, situational lexical functions, and that is linked to an ILU^c. Once these models and rules are formalized, we will initialize the population of the ILexicOn and the SLexicOn with concepts from other lexical ontologies.

Bibliography

Baker, C.F. and Fillmore, C.J. and Lowe, J.B., 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, 16-90*, ACL.

Boguslavsky I., 2002. Some Lexical Issues of UNL. *Proceedings of the First International Workshop on UNL, other interlinguas and their applications, Las Palmas*, 19-22.

Boguslavsky, I., 2005. Some controversial issues of UNL: Linguistic aspects. *Research on Computer Science*, 12:77-100.

Dong, Z. and Dong, Q. and ebrary, Inc, 2006. *HowNet and the Computation of Meaning*, World Scientific.

Gangemi, A. and Guarino, N. and Masolo, C. and Oltramari, A. and Schneider, L., 2002. *Sweetening ontologies with DOLCE*, Knowledge engineering and knowledge management: Ontologies and the semantic Web, 223-233, Springer.

Lenat, D.B., Guha, R.V., Pittman, K., Pratt, D., & Shepherd, M., 1990. Cyc: toward programs with common sense, *Communications of the ACM*, 33(8):30-49.

Mel'čuk I.A., Clas, A., Polguère, A., 1995. *Introduction à la lexicologie explicative et combinatoire.*

Mel'čuk I.A. 2004. Actants in semantics and syntax I: Actants in semantics. *Linguistics*. 42(1):1-66.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K.J., 1990. Introduction to wordnet: An on-line lexical database*. *International Journal of lexicography*, 3(4):235-344.

Niles, I. and Pease, A., 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, ACM.

Nirenburg, S. and Beale, S. and Mahesh, K. and Onyshkevych, B. and Raskin, V. and Viegas, E. and Wilks, Y. and Zajac, R., 1996. Lexicons in the Mikrokosmos project. In *Proceedings of the Society for Artificial Intelligence and Simulated Behavior Workshop on Multilinguality in the Lexicon, Brighton, UK*.

Polguère, A., 2009. Lexical systems: graph models of natural language lexicons. *Language resources and evaluation*. 43(1):41-55. Springer.

Sowa, J.F. 1984. *Conceptual structures: information processing in mind and machine*, System programming series, Addison-Wesley.

Vossen, P., 1998. EuroWordNet a multilingual database with lexical semantic networks, *Computational Linguistics*, 25(4).

XGR-CWL, 2008, Report of W3C Incubator Group on Common Web Language, `http://www.w3.org/2005/Incubator/cwl/XGR-cwl-20080331/`

Xenomarkers in Russian

Irina Levontina

Russian Language (Vinogradov) Institute, Moscow
irina.levontina@mail.ru

Abstract

Narrative retelling is usually considered within the frame of evidentiality (verification) that constitutes a grammatical category, such as special mood or similar, in some languages, for example, in American Indian, Tibeto-Burman, Bulgarian, Lithuanian, or Turkish. Cf. the use of Konjunktiv I in German: *Er habe das vergessen* (‘According to his words, he has forgotten’) by contrast with the indicative form: *Er hat das vergessen* (‘He has forgotten’).

As far as Russian is concerned, the same function is usually ascribed to the so-called “xenomarkers”, more specifically, to the particles *мол, дескать, де,* as well as *якобы*. Since the speaker uses xenomarkers to distance himself or herself from another person’s stand, these words quite often pragmatically imply a valuation, most often a negative one, of the reported speech.

It turns out, however, that the repertoire of means used as markers of quotation or retelling is much broader than it is generally admitted. Thus, the words *ах, вот, так и так*; the construction with imperative reduplication and the conjunction *да* (*Привязалась: расскажи да расскажи*), specific intonations of retelling, and some other phenomena can take over the same function.

Keywords

Xenomarkers, quotation, semantics, intonation, Russian language.

1 General remarks

Xenomarkers (or quotatives, or quotation and rendering markers) have been attracting linguists' attention long since. Thus, R. Jakobson in his classical work on shifters mentions the evidential mood in Bulgarian. Bulgarian has two forms, namely direct narration and indirect narration. Jakobson discusses a dialogue about a boat, where the form *zaminala* means 'it is claimed to have sailed' while *zamina* means 'I bear witness; it sailed' [Jakobson 1957]. It goes without saying, that similar meanings can be expressed in different languages. Note the use of Konjunktiv I in German: *Er habe das vergessen* ('According to his words, he has forgotten') by contrast with the indicative form: *Er hat das vergessen* ('He has forgotten').[1].

Narrative retelling is usually considered within the frame of evidentiality (verification) that constitutes a grammatical category, such as the special mood or similar, in some languages, for example, in American Indian, Tibeto-Burman, Bulgarian, Lithuanian, or Turkish. Cf. [Slobin, Aksu 1982; Chafe, Nichols 1986; Anderson 1986; Willett 1988; Bybee et al. 1994; De Haan 1998; Эвиденциальность 2007]. Compare also such terms as *quotative evidential, renarrative* [Плунгян 2008]. It should be stated however that in the above mentioned article a certain doubt is expressed that items in question are in fact semantically homogenous. Plungian suggests distinguishing between renarrative evidentiality markers on one hand, and subjective quotation modal markers (модализованные показатели «субъективного цитирования») on the other hand. Nevertheless we are not making this distinction here.

As far as Russian is concerned, the same function is usually ascribed to the so-called "xenomarkers", more specifically, to the particles *мол, дескать, де*, as well as *якобы* and *грит (гыт)*. Most of them are etymologically connected with verba dicendi. Since the speaker uses xenomarkers to distance himself or herself from another person's stand, these words quite often pragmatically imply a valuation, most often a negative one, of the reported speech. Special attention has been paid to *дескать* and *мол* [Отин 1966; Колодезнев 1969; Fontain

[1] It goes without saying that this idea can be expressed directly (*Он сказал, что…, По его словам…*). In writing quotation marks, in Russian spoken public speech the expressions, «Цитата» and «Конец цитаты» are used. Compare also the new in Russia iconic gesture "quotation marks", performed simultaneously with two fingers on each hand.

1983; Камю 1992; Баранов 1994; Арутюнова 2000; Шестухина 2003, Levin-Steinmann 1997].

These two particles can function in direct as well as in indirect or experienced speech. Their position in a phrase is relatively free. Their function, according to N. D. Arutiunova, consists in "marking Somebody Else's presence" («маркировать присутствие Другого») [Арутюнова 2000: 448]. It should be mentioned, that a speaker can present this way his or her own utterance, which took place earlier or is planned, as well as an interpretation of a person's non-verbal behavior.

In [Камю 1992] an attempt is made, to fix some semantic differences between the particles *мол* and *дескать*, although these distinctive features are subtle and are rather slight preferences. Roughly, *мол* is more closely connected with the original utterance, while *дескать* allows unrestricted interpretation of the situation in question.

The finest description of semantic differences between *мол, дескать*, and also *де* can be found in [Баранов 1994]. The meaning of these particles is considered within the framework of "the self" and "the other" («свой-чужой») opposition and the idea of "communicative responsibility": «*Дескать* отражает нежелание говорящего брать на себя ответственность за чужое (в целом или частично), а *мол*, напротив, свидетельствует о том, что за какие-то фрагменты чужого опыта он готов разделить ответственность с автором цитаты». [Баранов 1994: 116]. In [Плунгян 2008] *мол* is considered as an approximate renarration marker, *дескать* – as an interpretative renarration marker, and *де* implies a valuation or makes an ironical comment.

Returning to the repertoire of xenomarkers in Russian, it should be mentioned that the word *якобы* differs from *мол* and *дескать* in some more evident and important properties. *Якобы* is incompatible with direct speech, it presupposes the rendering of somebody's words rather de re, than de dicto. Moreover, *якобы* expresses a speaker's doubt about the contents of the reported utterance. See also [Плунгян 2008].

It turns out, however, that the repertoire of means used as markers of quotation or retelling is much broader than it is generally admitted.

2. *Ах*

Thus, the word *ах* can take over the same function, cf.:

- *Время идет быстро, а между тем здесь такая скука! - сказала она, не глядя на него.*
- *Это только принято говорить, что здесь скучно. Обыватель живет у себя где-нибудь в Белеве или Жиздре - и ему не скучно, а приедет сюда: "**Ах,** скучно! **ах,** пыль!" Подумаешь, что он из Гренады приехал.* (А. П. Чехов, Дама с собачкой).

"The days pass quickly, and yet one is so bored here," she said, not looking at him.

"It's the thing to say it's boring here. People never complain of boredom in godforsaken holes like Belyev or Zhizdra, but when they get here it's: '**Oh**, the dullness! **Oh**, the dust!' You'd think they'd come from Granada to say the least." (Anton Chekhov, Lady with Lapdog) [*As* The Lady with the Dog, *translated by Ivy Litvinov,* A. P. Chekhov: Short Novels and Stories, *Moscow: Foreign Languages Printing House, no date.*]

Note that the word *ax* cannot be used in a similar utterance outside of the context of retelling, cf.:

- *Вы хорошо съездили?* [Did you have a pleasant trip?]
- * *Ах, скучно! ах, пыль!* [Oh, the dullness! Oh, the dust!]

The central meaning of the interjection *ax* corresponds to a specific part of the human emotional spectrum (Cf. Tsvetaeva's «*Ах», когда чудно*). The context of the reported speech, however, loosens the restrictions on its use. Sentences with *ax* as a xenomarker may have two kinds of prosodic arrangement. *Ах* can function as a proclitic (*ахскуучно, ахпыыль*. A famous Russian reciter Dmitry Zhuravlev intones this phrase like this), at that, the word is prolonged, and the tone slightly rises and then falls down. *Ах* can also be pronounced separately, in this case with the rising tone, and the next word is pronounced with falling intonation (So does the other reciter of "Lady with Lapdog" Igor Yasulovitch).

3 *Вот*

One more interesting xenomarker was discovered in [Подлесская, Кибрик 2009], namely the interjection *вот* as a means of "expressing threat and condemnation" in reported speech, cf.:

\тоже на меня ' /посмотрела ,

«/\Вот!

Я тебя /–в-выгоню-у и-из-зз ··(0.2) этой из ш= ··(0.2) 'из ··(0.2) ' /школы!» [She looked at me too: "I'll throw you out of this school!"].

See also:

Я стала говорить дома: вот, Наташа просила у меня прощения, я не простила ее... [I started telling at home – well, Natasha asked me to forgive her, and I didn't…] [Н. Горланова, Метаморфозы].

Interestingly enough, the meaning of *вот* is not confined to threat or condemnation. Compare the following examples:

А она сидит и ноет: «Вот, я такая несчастная…» [And she is sitting around, whimpering: "Oh, I'm so miserable…"];

Он расхвастался: «Вот, я самый крутой» [He started bragging: "Yeah, I'm such a cool guy"];

Привязалась: «Вот, как тебе не стыдно, что у тебя за юбка» [She kept intruding on me: "Hey, what kind of a skirt is that? Shame on you!"];

А он все обещает: «Вот, деньги будут со дня на день, все отдам [And he keeps giving promises: "The money will be there any day, I will return everything I owe"]».

Ну и что же, что она первая позвонила? А ты бы ей сказал: «Вот, я сам собирался тебе позвонить, поздравить» [Yes, she called first, so what? You should have said – I was planning to call you with the greetings myself].

Now we are going to point out the common features of *ах* and *вот* as well as semantic differences between them. Both items function as a sort of forward quote, marking the beginning of reported speech. However, unlike *мол* and some other particles, these two items do not presuppose that the speaker is going to cite someone's speech. Normally they antecede a brief retelling of reported speech. The word *вот* tends to choose one or several representative phrases at that, while *ах* rather conveys the general idea of the speech and its emotional coloring. Moreover, *ах* stresses, that the speech in question was too emotional.

Russian has some more segmental xenomarkers at its disposal.

4 *Так и так*

Он пришёл в городское ГАИ и сказал, что вот так и так, Кио на гастролях, неизвестно, когда вернётся [He came to the local traffic police office and said that Kio is on tour and no one knows when he will be back]. [И. Э. Кио. Иллюзии без иллюзий (1995-1999)]

Прошло, наверное, уже полгода занятий в институте, когда он решился профессору сообщить: так и так, мол, мы с вами, Николай Васильевич, в некотором роде знакомы.

[The half of the academic year has passed when he ventured on telling the professor – Nikolaj Vasil'evich, you and I have, um, met before][Юрий Трифонов. Дом на набережной (1976)]
In [Плунгян 2008] this item is considered as an explicit marker of approximate retelling. In my opinion *так и так* rather involves the idea of well-ordered narration, step by step.

5 *Видите ли*

In sentences like *Он, видите ли, занят!* [He is supposed to be busy] a specific meaning of *видите ли* is presented, namely, the meaning of disapproving retelling. Note, that the "regular" *видите ли* has both plural and singular forms *(видишь ли / видите ли)*, while the "rendering" *видите ли* – is presented only by the plural.

6 Construction *расскажи да расскажи*

The construction with imperative reduplication and the conjunction *да* (*Привязалась: расскажи да расскажи [She is intruding on me: tell me! tell me!]*) also expresses the meaning of disapproving retelling. In this case the speaker does not like persistence and importunity of the author of reported speech. See also [Теоретические проблемы… 2010: 205]: *На меня стали бросаться: расскажи да расскажи им секрет самогона* [They started asking me to tell them the secret of making samogon] (И. Ильф, Е. Петров). Note that this construction cannot be used in a similar utterance outside of the context of retelling. Cf.:
Пристал, как с ножом к горлу: «Скажи да скажи про Аматурова Софье Михайловне: это скорей дело подвинет!» [He stuck to me like glue: tell Sofja Mihajlovna about Mr. Amaturov, it will help!] [А. Ф. Писемский. Просвещенное время (1875)]
Подсыпался к ней однажды бухгалтер: дай да дай выручку на два дня [The bookkeeper stuck to her once asking to lend him some cash from the box office for two days]. [Ю. О. Домбровский. Факультет ненужных вещей].

17 Intonation

In [Янко 2008: 109] an intonation pattern is described, called "the intonation of mental activity", i.e. situations of remembering, perplexity, sinking into daydreams, and also reported

speech (*Тетя сказала, надо чего-то там уко-олы делать [Aunty said we should make injections…]*). This intonation pattern is described as follows: «Соответствующий акцент характеризуется подъемом тона и существенным удлинением ударного слога акцентоносителя ремы. Вся заударная область ровная (иногда с небольшим естественным падением)».

Yanko notes, that by retelling a speaker doesn't copy somebody else's intonation, but arranges his or her utterance with a specific "remembering" prosody.

However, although the prosody of retelling and remembering has much in common, these two intonation patterns are somewhat different. First of all, retelling is often affective and emotive, and the prosody in this case is emphatic, which is hardly possible in the case of remembering, perplexity, or sinking into daydreams.

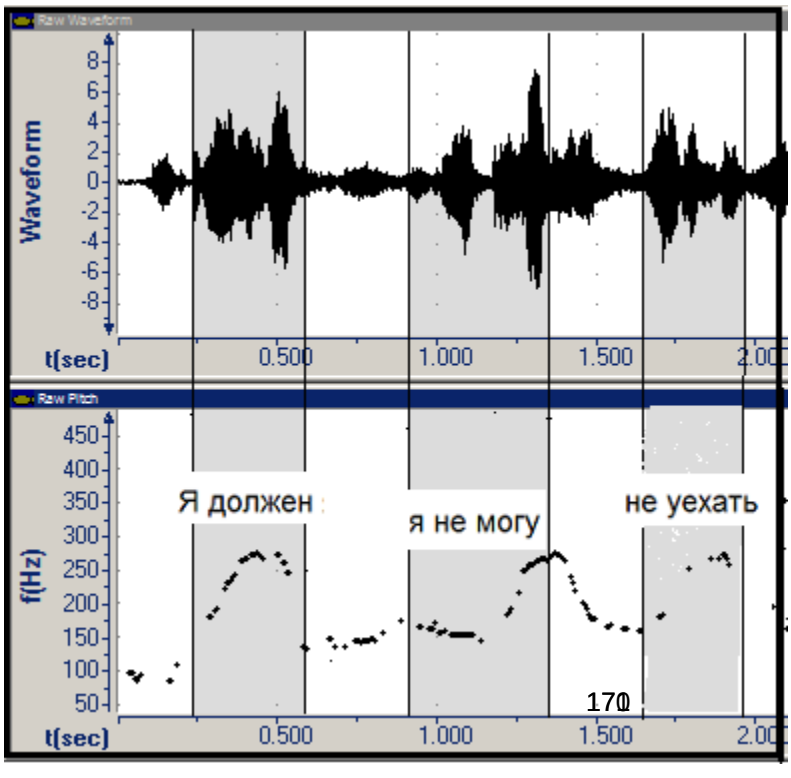
Secondly, by retelling, the phrase is often split into minor segments as against original speech. Cf.:

- *И что он ответил?*

- *Да что ответил! «/\Маама не разре/\шааает»*[2] [And what did he say? What could he say? Mommy won't let me].

From the point of view of intonation, reported speech often turns to sound rhythmical, pronounced with seriate tone rises and falls, similar to "listing" intonation: *А он мне и говорит: «/\Воот, /\деевушка, какая вы кра/\сиивая, как вас зо/\вут, а пой/\деемте, погу/\ляем, а /\даайте теле/\фоончик»* [And he tells me: Hey lady, you're so pretty, what's your name, let's go for a walk, please give me your phone number]. See the picture below:

Figure 1.



The phrase *(Ты говорил:) Я должен, я не могу не уехать* [(You said:) I must go, I cannot stay here] from the film «Проездом» (“Drive-by visit”) (1982).

“Retelling” intonation distorts the original utterance so that a listener understands, that the respective speech fragment doesn’t belong to the speaker. Thus, in the phrase *Как вас зовут?* [What is your name?] rising tone on *зовут* is in regular Russian impossible. But in reported speech slight tone rise on this word is quite natural.

Last but not least, if a speaker disapproves the original utterance, such phenomenon as “aping” (передразнивание) may appear. There are different means of aping in Russian. Now we will mention only two of them. In [Камю 1992] S. Kodzasov’s remark is adduced, that nasalization is often used as a means of aping.

One more means of this kind is “bleating” («блеяние») (*ма-а-а-ама*).

8 Speech substitutes

There is one more interesting phenomenon, connected with someone else’s speech reporting. Besides xenomarkers, there are some items, substituting or imitating someone’s speech. Usually such items are senseless combinations of sounds, including iterations and rhymes and characterized with the same intonation pattern as was already discussed. These are units like *ля-ля тополя, ля-ля-фа-фа, тэ-тэ-тэ, тэ-тэ-нэ-нэ, тэто-это, тра-ля ля* and a relatively new borrowing *бла-бла(-бла)*; *Я ему объясняю: «У меня много работы, а завтра теща приезжает, тэ-тэ-нэ-нэ <ля-ля тополя>…»* [I’m telling him: I’ve got a lot of work, and tomorrow my mother-in-law is coming, bla-bla-bla]; *Ты ему скажи, что ты к нему хорошо относишься, но только как к другу, бла-бла-бла* [You should tell him that you like him but only as a friend, bla-bla-bla]. Compare also: *Прибегает: «А! О!» А чем я могу ему помочь?* [He came crying “Ah! Oh!”, but how could I help him?]; *Опять наехала на меня: «Аа! Даа!» Надоела уже* [She started picking on me again: “Aa! Oo!” – I’m sick of her]. Moreover there are expressions *тыры-пыры* and *тыр-пыр восемь дыр*. Normally such units are not reflected in written texts, but quite often occur in oral discourse. Some of these expression are used not only to substitute someone else’s speech, but also to denote it: *Вот сейчас, я пролистал ЖЖ и думаю что-то вроде - вот я лошара и неудачник, так бездарно проводил время и сох па разным тёлкам и всё бес толку и писал какую-то пургу какие-то рассказы и бла бла*

[Just now I looked up my LJ and I'm thinking something like 'I'm such a jerk and a looser, I wasted so much time dreaming of different babes, writing all kinds of fluff and bla-bla-bla].

It should be mentioned that some items with a broader meaning also often function as xenomarkers. Thus, the word *вроде* – a marker of uncertainty – can be used in reported speech as xenomarker. Compare two sentences: *Ты вроде похудела* [You seem slimmer] [the speaker is not sure] and *Он вроде уволился* [They say he quit the job]. The word *типа* – a marker of approximate nomination – can function in a similar way: *А она стала говорить, что муж типа так занят* [And she started saying that her husband is sorta busy]. As was already noticed (see fox example [Булыгина/Шмелёв 1993; 1997, Арутюнова 1999]), this meaning is typical of some expressions with the verb *говорить* ('to say'), predicatives *слышно*, *похоже*, etc., and also constructions with (*как*) *будто* (*бы*) and some others. In [Летучий 2008] comparative constructions with conjunctions *как будто*, *будто*, *словно*, *как бы*, *как будто бы*, *будто бы* are considered in detail, as a source of units with the meaning of evidentiality.

One more interesting item is a new meaning of the pronominal *такой*: *А я такая: «Как тебя зовут?»* [And I'm like – what's your name?]. Sometimes it is considered as an evidential marker, because very often it precedes direct speech [Савчук 2011]. In my opinion it has a rather iconic function (a speaker wants the listener to vividly imagine what happened) and does not obligatory presuppose reported speech: *А я такая подхожу, беру сигарету и закуриваю. Все в шоке* [And I'm like coming up, taking a cigarette and lighting it. All are shocked].

So we tried to show that xenomarkers in Russian are numerous, various and belong to different levels of language. They demand careful examination.

Bibliography

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З. *Теоретические проблемы русского синтаксиса. Взаимодействие грамматики и словаря*, Москва: Языки славянских культур, 2010.

Арутюнова Н.Д. 1999. *Язык и мир человека*. М.: Языки русской культуры.

Арутюнова Н.Д. 2000. Показатели чужой речи *де, дескать, мол* //*Язык о языке*. Под общ. рук. и ред. Н.Д. Арутюновой. Москва: С.437-452.

Баранов А. Н. 1994. Заметки о дескать и мол // *ВЯ*: № 4, 114-124.

Булыгина Т.В., Шмелёв, А.Д. 1993. Гипотеза как мыслительный и речевой акт // Н. Д. Арутюнова (ред.). *Логический анализ языка: Ментальные действия*. М.: Наука, 78-82.

Булыгина, Т.В.; Шмелёв, А.Д. 1997. Гипотезы и квазиассерции // Т.В. Булыгина; А.Д. Шмелёв. *Языковая концептуализация мира (на материале русской грамматики)*. М., Школа «Языки русской культуры», 293-304.

Камю Р. 1992. *Мол, дескать, -де*: чужая речь в чужом языке // *Проблемы интенсивного обучения неродным языкам (материалы первой международной научно-методической конференции, 27-28 мая 1992 г.)*, Российский государственный педагогический университет имени А.И. Герцена, Санкт-Петербург: «Образование», 52-57.

Колодезнев В. М. 1969. О значении частиц *мол, де, дескать* // *Русский язык в школе*: № 1.

Летучий, А.Б. 2008. Конструкции сравнения ситуаций с показателями *как бы* и *как будто* // *Wiener Slawistischer Almanach*, Sonderband 72. München: Sagner.

Отин Е. С. 1966. О субъективных формах передачи чужой речи // *Русский язык в школе*: № 1.

Плунгян В. А. 2008. О показателях чужой речи и недостоверности в русском языке: *мол*, *якобы* и другие // Wiemer B. & V. A. Plungjan (Hrsg.). *Lexikalische Evidenzialitäts-Marker in slavischen Sprachen* (Wiener Slawistischer Almanach, Sonderband 72.) München: Sagner, 285-311.

Подлесская В. И., Кибрик А. А. 2009. Дискурсивные маркеры в структуре устного рассказа: опыт корпусного исследования // *Диалог*.

Разлогова, Е.Э. 1996. Модальные слова и оценка степени достоверности высказывания // *Русистика сегодня* № 3, 21-47.

Савчук С. О. 2011. Местоимение такой в функции маркера чужой речи в устном высказывании // *Вопросы культуры речи*, вып. 10.

Шестухина И. Ю. 2003. Эмоционально-экспрессивное и функционально-стилистическое значение частиц *мол, де, дескать* в русском языке // *Мат. Всероссийской научно-практ. конф. «Русский язык и культура речи как дисциплина государственных образовательных стандартов высшего профессионального образования: опыт, проблемы, перспективы».* Барнаул: Изд-во АГУ, 366-368.

Эвиденциальность в языках Европы и Азии. Сборник статей памяти Наталии Андреевны Козинцевой 2007. М.: Наука.

Якобсон Р.О. 1972. Шифтеры, глагольные категории и русский глагол // *Принципы типологического анализа языков различного строя*. М.

Янко Т. Е. 2008. *Интонационные стратегии русской речи в сопоставительном аспекте*. М.

Aikhenvald A. 2004. *Evidentiality*. Oxford: Oxford U. P.

Anderson L. 1986. Evidentials, paths of change, and mental maps: typologically regular asymmetries. // Chafe & Nichols (eds.), 273-312.

Bybee J., R. Perkins, W. Pagliuca 1994. *The evolution of grammar: Tense, aspect and modality in the lanuages of the world*. Chicago: University of Chicago Press.

Chafe W., J. Nichols (eds.) 1986. *Evidentiality: the linguistic coding of epistemology*. Norwood: Ablex, 1986.

De Haan Ferdinand 1998. *The Category of Evidentiality*. Ms.

Fontain J. 1983. *Grammaire du texte et aspect du verbe russe contemporain*. P.

Jakobson R. 1957. "Shifters, verbal categories, and the Russian verb," in *Selected Writings*, vol. II, *Word and Language*, The Hague: Mouton, 1971.

Levin-Steinmann, A. 1997. *Мол, дескать* und *де* – alles ein und dasselbe, oder doch nicht ganz? // J. Schulze & E.Werner (Hrsg.), *Linguistische Beiträge zur Slavistik*. München: Sagner, 207-231.

Plungian, V.A. 2001. The place of evidentiality within the universal grammatical space // *Journal of Prag-matics* 33.3, 349-357.

Rakhilina, E.V. 1996. *Jakoby* comme un moyen de médiatisation en russe // Z. Guentchéva (ed.). *L'énon-ciation médiatisée*. Paris: Peeters, 299-304.

Slobin D., and A. Aksu 1982. Tense, Aspect, and Modality in the use of the Turkish evidential // In Hopper P.J. (ed.) *Tense-aspect: Between Semantics and Pragmatics*. Amsterdam: Benjamins, 185-200.

Willett T. (1988). A cross-linguistic survey of grammaticization of evidentiality // *Studies in Language*: 12.1, 57-91.

Assessing and Improving Paraphrasing Competence in FSL

Jasmina Milićević and Alexandra Tsedryk

Dalhousie University
6135 University Avenue
Halifax, Nova Scotia, B3H 4P9, Canada
jmilicev@dal.ca | atsedryk@dal.ca

Abstract

We present preliminary results of a study whose goals are 1) to test the paraphrasing competence of advanced adult learners of French as a Second Language (= FSL), 2) assess their learning needs and 3) suggest possible tools for teaching them paraphrasing techniques.

Keywords

Paraphrasing competence, adult FSL learner, user-friendly paraphrasing rules

1 Paraphrasing Competence in the First and Second Language

As noted by many (for instance, Žolkovskij & Mel'čuk 1967, Fuchs 1980, Mel'čuk 1992, Martinot 2003), paraphrasing competence, i.e., the capacity to produce paraphrases, or quasi-synonymous sentences, like those in (1), is part and parcel of linguistic competence.

(1) a. *Marc a toujours beaucoup de questions pour le professeur de français* 'Marc has always many questions for the French teacher.'

b. *Marc pose toujours beaucoup de questions dans le cours de français* 'Marc asks always many questions in French class.'

Native speakers (henceforth, L1 speakers) need paraphrasing in order to get around difficulties inherent to speech production (restricted lexical and syntactic co-occurrence, lexical gaps, etc.) or to reformulate their discourse for reasons of clarity and style. This is all the more true for language learners (or L2 speakers), who desperately need spare paraphrastic variants to avoid "crashes in generation". (Paraphrases are also needed for speech comprehension, as a way to seize the content of an utterance by reformulating it. However, in this paper we will set aside the comprehension of paraphrases and concentrate solely on their production.)

Recently, there has been growing interest in the application of linguistic models of paraphrasing in language assessment and teaching; although not to the extent one would hope for, considering the importance of this phenomenon.

Assessment of the paraphrasing competence in French speaking children aged 4 to 10 is the subject of Martinot (in press). Paraphrasing competence is acquired gradually, throughout childhood, starting with the acquisition of (in Martinot's terms) descriptive paraphrases (*Elle sera ta voisine* 'She [a classmate] will be your neighbor' ~ *Elle sera à côté de toi* 'She will be [sitting] at

your side'), followed by semantic paraphrases (*Julie chuchota à Tom* … 'Julie whispered to Tom' ~ *La petite souffla à son voisin* … 'The little girl murmured to her neighbor') and formal paraphrases (*Elle tenait par la main une petite fille* 'She was holding by hand a little girl' ~ *La maîtresse, elle la tient par la main* 'The teacher, she is holding her by hand') till, finally, extralinguistic paraphrases are acquired (*Un jour une nouvelle fille arriva dans une école* 'One day a new girl came to a school' ~ *Elle tenait par la main une petite fille que personne n'avait encore jamais vue* 'She was holding by hand a little girl whom no one had seen before').

Russo & Pippa (2004) devised a test in which the ability to paraphrase was used as a predictor of the ability to interpret. Students of an interpretation school who scored high on the paraphrase test were also more successful in interpreting and graduated sooner than those who had poorer paraphrasing skills. They resorted to complex paraphrasing techniques, reformulative paraphrase in the authors' terms, making use of implication, condensation, generalization and particularization of meaning. The following is an example of condensation: […] *capace di apportare una soluzione di pace e guistizia in Medio Oriente* 'able to bring a solution of peace and justice to problems in the Middle East' ~ […] *per trovare una soluzione ai problemi medioorientali* 'in order to find a solution for Middle East problems'.

Paraphrasing for specific purposes—as a way to avoid plagiarism—was studied, for example, in Keck (2006) and McInnis (2009). The former distinguished four types of reformulations, based on their proximity to the original text: near copy, minimal revision, moderate revision and important revision. An example of the last type of modification follows: *Children speak more like adults, dress more like adults and behave more like adults than they used to*. ~ *It seems that things that children do and even what they wear are more adult-like than ever before*. McInnis (2009) reports that L1 and L2 speakers alike have insufficient skills when it comes to reformulation and insist on the importance of teaching paraphrasing techniques in academic writing courses.

As for teaching tools intended to boost paraphrasing competence of language learners, their development has received even less attention. Polguère (2004) and Milićević (2008) and (2009) propose learner-friendly adaptations of linguistic formalisms used by Meaning-Text theory (Mel'čuk 1997, Kahane 2003) for paraphrase modeling: lexical functions (Wanner, ed., 1996) and paraphrasing rules (Žolkovskij & Mel'čuk 1967, Mel'čuk 1992, Milićević 2007).

In the rest of this paper we present our own test for assessing the paraphrasing competence of FSL learners (section **2**) and offer some suggestions as to how to improve the latter (section **3**). We conclude with a brief summary of the research presented in the paper (section **4**).

We target adult FSL learners and advocate an explicit teaching of paraphrase and related concepts. Our approach is anchored in Meaning-Text linguistic theory, a framework that has placed paraphrasing in the centre of linguistic research and has developed powerful formal tools for its modeling, tools that can be—and, as we saw, have to some extent already been—adapted for teaching purposes.

2 Assessing Paraphrasing Competence of FSL Learners

We now describe the test and present a qualitative and quantitative analyses of the results.

2.1 Test Design

We asked some twenty FSL learners, first and third year university students having gone through French immersion prior to university, to propose paraphrases for five sentences—sentence (1) and the following four:

(2) a. *Nathalie a correctement analysé cette phrase difficile* 'Nathalie correctly analyzed that difficult sentence.'

b. *Julie admire énormément l'auteur de ce roman historique* 'Julie admires enormously the author of this historic novel.'

c. *Paul a beaucoup d'enthousiasme pour le théâtre* 'Paul has a lot of enthusiasm for the theatre.'[1]

d. *Ce professeur m'a donné un bon conseil* 'This professor gave me a good piece of advice.'

These sentences were chosen because, while being lexicaly and structurally relatively simple, they could in principle yield a large number of paraphrases, involving, in particular, nominalizations and support verbs, i.e., transformations the students participating in the test were supposed to be familiar with.

A control group consisting of an equal number of native speakers performed the same task, the hypothesis being that the natives would perform better. Neither group had benefited from special training concerning paraphrasing; the concept was explained and illustrated immediately before the test.

The participants were supposed to give three paraphrases per sentence (some produced less). Sentences judged not sufficiently synonymous to the initial ones were discarded; for instance, (3a) was not recognized as a valid paraphrase of the sentences in (1). Sentences that preserved the initial meaning but presented errors or stylistic inadequacies were retained as "defective" paraphrases ; one such paraphrase of (1) is (3b), which means the same as the former but contains a wrong support verb for QUESTION 'question' (**demander* lit. 'ask' instead of *poser* lit. 'put').

(3) a. *La curiosité de Marc pour le français est sans fond* 'Marc's curiosity for French is without limit.'

b. *Marc *demande toujours beaucoup de questions à son professeur de français* 'Marc always asks a lot of questions to his French teacher.'

Following this method, we obtained two corpora of paraphrases, one produced by L2 and the other by L1 speakers, each containing some 300 sentences.

We compared the two corpora looking for significative differences with respect to 1) the preservation of the paraphrasing link (is the proposed sentence a valid paraphrase of, i.e., sufficiently synonymous with, the initial sentence?); 2) the exactness of the paraphrasing link (is the proposed sentence an exact or near paraphrase of the initial sentence?); 3) the paraphrasing techniques used (inference, semantic decomposition, lexical substitution…) and 4) the formal correctness (grammaticality and stylistic acceptability) of the proposed paraphrases.

2.2 Qualitative Analysis of Results

In both L1 and L2 corpora there were only a few cases in which the paraphrasing link with the initial sentence was not preserved. The vast majority of valid paraphrases in both corpora were approximate (rather than exact) paraphrases. As for the types of paraphrases represented in our corpora, we refer to them according to the standard Meaning-Text paraphrase typology

[1] This sentence is ambiguous, due to the polysemy of THÉÂTRE 'theater', which can be interpreted either as 'theater plays (one is watching)' or 'artistic activity (one is involved as an actor)'. The first reading is pragmatically more plausible, but we got paraphrases for both (which, of course, are not synonymous); cf., respectively, *Paul aime regarder les pièces de théâtre* 'Paul likes watching theater plays' and *Paul aime faire du théâtre* 'Paul likes doing theater'.

(Milićević 2007: 138*ff*). Both corpora contained a full range of paraphrase types distinguished within our framework, albeit used in different proportions (see below, subsection **2.3**).

- **Types of paraphrases in L1 & L2 corpora**

Extralinguistic paraphrases

Roughly speaking, these are paraphrases whose production requires, beside the knowledge of a language, some encyclopedic and pragmatic knowledge, as well as the use of logical capabilities. In our corpora, extralinguistic paraphrases came in three subtypes—referential, situational and encyclopedic, cf., respectively:

(4) a. *Nathalie* 'Nathalie' ~ *l'étudiante* 'the student'; *le* [professeur] 'the [teacher]' ~ *ce/son* [professeur] ~ 'this/his [teacher]'

b. [professeur] *de français* '[teacher] of French' ~ [professeur] *de langue maternelle* '[teacher] of mother tongue'

c. [professeur] *de français* '[teacher] of French' ~ [professeur] *qui enseigne la langue de Molière* '[teacher] who teaches the language of Molière'[2]

Linguistic paraphrases

We outline here four major linguistic paraphrase types and only some of their subtypes, without mentioning the cases in which these different types were combined to produce a pair of paraphrases (which actually happened quite often).

—Semantic paraphrases

These paraphrases are either propositional (i.e., concerning a modification or a different expression of the propositional meaning), further divided into inferences, replacements, additions and decompositions, as in (5a)-(5d), or communicative (concerning different ''information packaging''), featuring theme and rheme focalization: see (5e)-(5f); in addition, (5e) has a different theme with respect to the source sentence.[3] [In the examples below, P stands for a lexical meaning corresponding to a predicate (in the logical sense).]

(5) a. *analyser P* 'analyze P' ~ *comprendre P* 'understand P'

b. *faire P toujours* 'still do P' ~ *ne pas cesser de faire P* 'not stop doing P'

c. *avoir analysé P* 'have analyzed P' ~ *avoir pu/su analyser P* 'have known how/have been able to analyze P'

d. *donner un conseil à quelqu'un* 'give a piece of advice to someone' ~ *lui dire ce qu'il doit faire* 'tell him what he should do'

e. [with respect to (2a)] *Cette phrase difficile, Nathalie l'a bien analysée* 'That difficult sentence, N. analyzed it well.'

f. [with respect to (2a)] *C'est cette phrase difficile que Nathalie a bien analysée* 'It is that difficult sentence that N. analyzed well.'

—Lexical-syntactic paraphrases

We illustrate only major paraphrase subtypes: synonymic substitutions, simple and with light verb fission (6a-b); antonymic substitutions (6c) and conversive substitutions, also simple (6d) and with light verb fission (6e). In the examples that follow, we use lexical

---

[2] It is quite possible that the phrase *langue de Molière* is so well-known to French speakers that is has to be considered as a synonym for *langue française*, in which case this would be an example of linguistic, more specifically, lexical paraphrases.

[3] We believe that some communicative paraphrases in the corpus were produced spuriously, i.e., as a non intended result of word-order modifications. However, we have no way of ascertaining this.

function symbols (in Monaco font) without any explanation; we are relying on the reader's intuitive understanding of these examples.

(6) a. [with Gener] *phrase* 'sentence' ~ *expression* 'expression'; [with Syn] *difficile* 'difficult' ~ *dure* 'hard'; [with Sing] *théâtre* 'theater' ~ *pièces de théâtre* 'theater plays'

b. [with S_0] *X admire Y* 'X admires Y' ~ *X éprouve de l'admiration pour Y* 'X feels admiration for Y'; [with A_1] *X a de l'enthousiame pour Y* 'X has enthusiasm for Y' ~ *X is enthousiaste concernant Y* 'X is enthusiastic when it comes to Y'

c. *difficile* 'difficult' ~ *pas facile* <*évident*> 'not easy <obvious>'

d. [$Oper_1$ ~ $Oper_3$] *X donne un conseil à Y* 'X gives a piece of advice Y' ~ *Y reçoit un conseil de X* 'Y gets a piece of advice from X'

e. [with S_2] *X admire Y* 'X admires Y' ~ *Y est un idole de X* 'Y is X's idol'; [with copula] *X admire Y beaucoup* 'X admires Y a lot' ~ *L'admiration de X pour Y est grande* 'X's admiration for Y is big'

—Syntactic paraphrases

Four major types of syntactic paraphrases were found: part of speech conversion, word order variation, passivization and restructuring; cf., respectively:

(7) a. *correctement* 'correctly' ~ *de façon correcte* 'in a correct manner'; *difficile* 'difficult' ~ *qui était difficile* 'that was difficult'

b. *a correctement analysé P* 'has correctly analyzed P' ~ *a analysé P correctement* 'has analyzed P correctly'

c. [with respect to (2d)] *Un bon conseil m'a été donné par ce prof* 'A good piece of advice was given to me by this teacher'

d. [with respect to (2a)] *Cette phrase était difficile mais Nathalie l'a bien analysée* 'That sentence was difficult but Nathalie analyzed it well.'

• Formally incorrect paraphrases in L2 corpus

The ungrammatical or stylistically unacceptable paraphrases, due to lexical or/and syntactic errors, were, of course, overwhelmingly present in L2 corpus.

Paraphrases featuring lexical errors

Lexically deficient paraphrases contained both incorrect derivations and incorrect collocates, as shown respectively in (8a) and (8b):

(8) a. [incorrect S_0] *X analyse Y correctement* 'X analyzes Y correctly' ~ **L'**analysation** de Y par X est correcte* 'X's analyzation of Y is correct'; [incorrect A_1] *X a de l'enthousiasme pour Y* 'X has enthusiasm for Y' ~ **X est **enthousiastique** concernant Y* 'X is enthusiast about Y'

b. [incorrect value of $Oper_1$] *X a beaucoup de questions for Y* 'X has a lot of questions for Y' ~ **X **demande** beaucoup de questions à Y* 'X asks a lot of questions to Y'; [incorrect value of Bon] *donner un bon conseil* 'give a good piece of advice' ~ **conseiller **soigneusement*** 'advise carefully'

Paraphrases featuring syntactic errors

Syntactically problematic paraphrases contained errors in linear ordering, passivization, relativization and government:

(9) a. [with respect to (2a)] ****Correctement,** Nathalie a analysé cette phrase difficile* 'Correctly, Nathalie has analyzed that difficult sentence'.

b. **Le professeur de français est toujours posé beaucoup de questions par Marc.* 'The French teacher is always asked a lot of questions by Marc'.

c. [with respect to (2b)] **L'auteur de ce roman est celui **qui** Julie admire énormément* 'The author of this historic novel is the one who Julie admires enormously'.

d. **aimer **d'**aller au théâtre* 'to like that goes to the theater'; **sans **faisant** d'erreur* 'without make mistake'

Quite a few errors in the L2 corpus can be attributed to negative transfer from English, which is the mother tongue of most of the FSL students who participated in the test. Cf., for instance, the form of the adjective in (8a) and the attempt to use the indirect passive (featuring the Adressee in the Subject position), inexistant in French but common in English, in (9b).

2.3 Quantitative Analysis of Results

In order to measure the preservation of paraphrastic links and formal correctness of paraphrases, all produced sentences were divided into three subsets: 1) **P**: sentences that preserved the initial meaning and contained no errors (cf. (1)); 2) **?P**: sentences that preserved the initial meaning but contained grammatical, lexical or stylistic errors (cf. (3b) with respect to (1)); 3) **nonP**: sentences that did not preserve enough the initial meaning (cf. (3a) with respect to (1)). The percentage of each type of sentences in the L1 and L2 corpora is given in Figure 1:

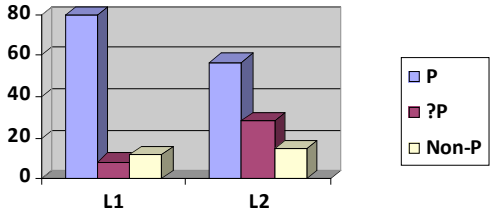


Figure 1: Preservation of paraphrastic links/Formal correctness of paraphrases: L1 *vs.* L2

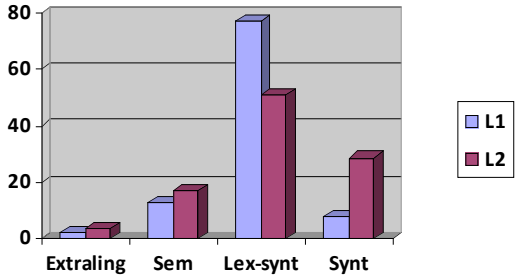


Figure 2: Paraphrase types (use of paraphrastic techniques): L1 *vs.* L2

Native speakers proposed more valid paraphrases than L2 learners (79, 4% *vs.* 56, 9 %), who in turn produced more formally incorrect paraphrases. The percentage of semantically equivalent sentences was relatively high in both groups while the percentage of non-paraphrases was low and approximately the same (12% for L1 and 15% for L2). These results indicate that both L1 and L2 participants intuitively understood the concept of paraphrase and were able to produce paraphrases.

The biggest difference observed is that the number of formally incorrect paraphrases was three times higher in the L2 corpus (28.1 %) than in the L1 corpus (8.6%).

With regard to paraphrastic techniques used by both groups of participants, we wanted to see, in particular, whether one group preferred certain paraphrasing techniques over others, and if so, in what proportion. We wanted to observe specific difficulties the L2 group experienced with paraphrastic techniques in order to be able to address them later on. Figure 2 shows the distribution (in percentages) of the four previously described types of paraphrases in the L1 and L2 corpora.

We can make the following three observations. First, both L2 and L1 subjects used all types of paraphrases. (This shows that there is no specific type of paraphrase completely ignored by L1 or L2 group.) Second, the use of extralinguistic paraphrases was rather restricted in both groups; they were mostly found in the L2 corpus (3.6 % in the L2 corpus *vs.* 2.2% in the L1 corpus). Third, the distribution of linguistic paraphrases was more or less even for the semantic type (12.5% in the L1 corpus *vs.* 16.8 % in the L2 corpus). It should be noted, however, that out of the 16.8% of semantic paraphrases in the L2 corpus, 5 %

represented communicative changes of the type illustrated in (5e)-(5f), which necessarily involve word order manipulations. The fact that this paraphrastic technique was found only in the L2 corpus indicates that L2 learners preferred syntactic restructuring of sentences to lexical substitutions and that, moreover, some of this restructuring may have been unintended (cf. footnote 3, p. 2). A more significant difference was observed in two types of paraphrases: lexical-syntactic and syntactic. The L1 subjects used more lexical means (77.5 % in the L1 corpus *vs.* 51% in the L2 corpus) and the L2 subjects produced more syntactic paraphrases (28.6 % in the L2 corpus *vs.* 7.8 % in the L1 corpus).

When it comes to the richness of expression observed in each group of subjects, L1 speakers were by far better able to vary their expression. Thus, in case of sentence (2a), they came up with 19 different lexical-syntactic reformulations, while L2 speakers provided only 9. For example, L1 subjects used five different substitutions for the adverb *correctement* 'correctly' (*brillamment* 'briliantly', *parfaitement* 'perfectly', *efficacement* 'efficiently', *bien* 'well', *très bien* 'very well') and L2 subjects only one (*bien* 'well'). Figure 3 provides details about the variability of expressions per participant in each group of subjects.

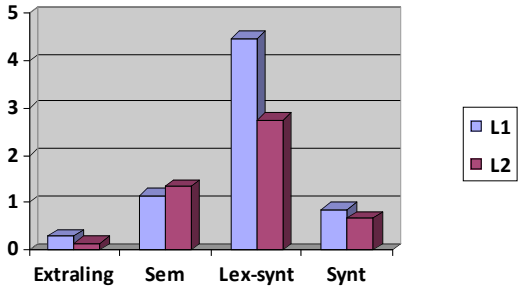


Figure 3: Variation of expression: L1 *vs*. L2

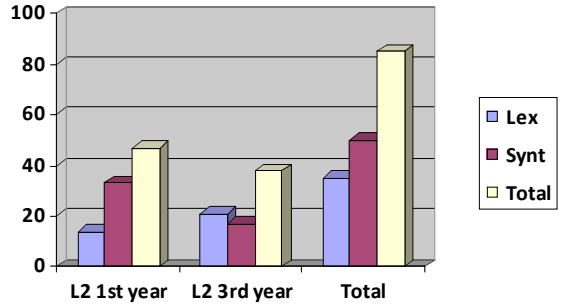


Figure 4: Error types in L2 (1st year *vs*. 3rd year)

The L1 and L2 groups differed most significantly in that the former proposed much richer lexical-syntactic variation per participant than the latter (4.48 *vs*. 2.76). L2 participants lacked lexical means and used less complex lexical-syntactic substitutions. Clearly, this is an area where there is room for improvement for the L2 group.

As far as formal correctness of paraphrases in the L2 corpus is concerned, we wanted to determine what types of errors, syntactic or lexical, were prevalent in the first as opposed to the third year of study. Figure 4 shows the proportion of lexical and syntactic errors identified in the first and third year students' sentences. (Sentences treated as "nonP" did not count for error analysis.)

Overall, L2 subjects made less lexical than syntactic errors. For first year students, the number of syntactic errors per participant was 2.54 and the number of lexical errors 1.07. We attribute this to the fact that their vocabulary is less developed and that because of this they relied mostly on syntactic paraphrasing means. Consequently, they were more likely to commit errors of this kind. On the other hand, with third year students we observe the opposite. Syntactic errors were less numerous than lexical ones (2.12 *vs*. 2.63 errors per participant). Third year students made twice as many lexical errors than first year students. This may be due to the fact that more advanced L2 learners are more confident in their vocabulary knowledge and, as a result, make more use of various lexical substitutions—also making more mistakes in the process. When trying to use synonymic substitutions with light verb fission, students often literally translated light verbs from English into French; 86% of all incorrect collocates, as in the example (8b), contained an inappropriate light verb.

Although our L2 corpus represents a relatively small sample, we observed an improvement with respect to the grammar of the third year learners. These observations corroborate some previous findings concerning L2 advanced learners: while possessing solid grammatical skills, these learners still produce lexical and stylistic errors (cf., for instance, Thomas 2008).

2.4 Summary of the Results

Preliminary results of the test corroborate our initial hypothesis, namely that the paraphrasing competence of native speakers is superior to that of non-natives, and give us other useful information, in particular:

1) The concept of paraphrase itself does not seem to be problematic for either group of subjects; in both cases, the number of sentences that were discarded as non-paraphrases (of the initial sentences) was insignificant.

2) The most frequent paraphrase types were in both cases semantic paraphrases (using semantic decompositions or inferencing) and lexical-syntactic paraphrases (using nominalizations, light verbs, etc.); both the natives and the non-natives preferred near-paraphrases (as opposed to exact ones).

3) The most striking difference between the two groups concerned the lexical and syntactic means used in paraphrase production: the francophones used a rich variety of (near-)synonyms and rather complex syntactic constructions; in contrast, quite a few FSL learners limited themselves to more or less local syntactic variation, exploiting only word order and passivization.

4) The difficulties of L2 learners can be explained by their insufficient lexical knowledge; this concerns both lexical relations and properties of individual lexical units, such as the goverment. Many errors were due to negative transfer effects.

3 For a Better Paraphrasing Competence of FSL Learners

Since the main reason for the inferior paraphrasing skills of L2 learners turns out to be their insufficient lexical knowledge, an obvious conclusion is that they need the instruction emphasizing this aspect of their L2. Let us give just one example (for simplicity's sake, in English) of how lexical and paraphrastic relations could be presented to language learners.

Suppose we want to teach our students lexical and paraphrastic relations between the verb (to) INTEREST (as in *Global warming interests scientists more and more*) and other lexical items of English.

We would start by presenting the propositional form of the verb in question, i.e., an expression featuring the verb itself and its semantic actants, X and Y, with the indication of their semantic class (PHENOMENON and PERSON, respectively):

(10) PHENOMENON *X interests* PERSON *Y*

Then we would give a number of paraphrases of the verb, based on lexical relations of synonymy, nominalisation, adjectivalization, etc.:

(11) a. *X* ***intrigues*** *Y*
b. *X* ***is interesting*** *for Y*
c. *X* ***is of interest*** *for Y*
d. *X* ***awakens*** <***arouses***> *Y's* ***interest***
e. *Y* ***is interested*** *in X*
f. *Y* ***shows interest*** *for X*

Table 1 shows the lexical relations involved in (11), in both the standard MTT and a learner-friendly form (shaded); paraphrasing rules—in both formats—necessary to produce paraphrases in (11) are given in Table 2.

Syn	synonymous V	(11a)
A_1	characteristics of X	(11b)
A_2	characteristics of Y	(11e)
S_0	N(V)	(11c/d/f)
$Oper_1$	[subject X] light V	(11b/c/e)
$Oper_2$	[subject Y] light V	(11f)
$Caus_1Func_0$	[subject X] causative V	(11d)

Table 1: Some lexical relations in the standard MTT and in a learner-friendly notation

$L_{(V)} \approx$	$Syn(L_{(V)})$	(11a)
$V_{[X \sim Y]} \approx$	synonoymous V	
$L_{(V)} \approx$	$Oper_1$—II→$A_1(L_{(V)})$	(11b)
$V_{[X \sim Y]} \approx$	BE + characteristics of X	
$L_{(V)} \approx$	$Oper_1$—II→$A_2(L_{(V)})$	(11e)
$V_{[X \sim Y]} \approx$	BE + characteristics of Y	
$L_{(V)} \approx$	$Oper_1$—II→$S_0(L_{(V)})$	(11c)
$V_{[X \sim Y]} \approx$	[subject X] light V + N(V)	
$L_{(V)} \approx$	$Oper_2$—II→$S_0(L_{(V)})$	(11f)
$V_{[X \sim Y]} \approx$	[subject Y] light V +N(V)	
$L_{(V)} \approx$	$Syn(L_{(V)})$	(11d)
$L_{(V)} \approx$	synonymous V	

Table 2: Some paraphrasing rules in the standard MTT and in a learner-friendly notation

Students are not obliged to see the information encoded in the standard way (although in some cases and for some types of learners this too could be useful): we provide it here only for the purpose of comparison.

Finally, we could point out a generalization, namely that for any verb belonging to the same semantic class as the verb (to) INTEREST, i.e., a causative attitudinal verb (some others: *astonish*, *disappoint*, *irritate*, etc.), one can in principle have these same paraphrase types.

Of course, the teaching techniques suggested above presuppose that quite a few lingustics notions must be taught at the same time. This is why they are well suited for adult learners who have had quite a few years of formal schooling. For other profiles of learners (less educated, younger ones, ...) less explicit methods would probably be better indicated.

Let us conclude by the following three observations. First, we believe that making students aware of the systematic and cross-linguistic nature of lexical and paraprhastic relations (and the corresponding formalisms) can help improve their language manipulation abilities. Second, this kind of instruction is good also for a contrastive analysis of learners' first and second languages and could help alleviate at least some negative transfer problems. Last but not least, the method proposed here for teaching about paraphrases in FSL is exportable, i.e., can be used in case of other languages as well.

4 Conclusion

In this paper we described a test designed to assess the paraphrasing competence of adult university-level learners of French and suggested possible ways of reinforcing it.

In order to provide this group of students with a proper training in the use of paraphrasing techniques, we believe it necessary to: 1) insist on the importance of paraphrases in language production—via explicit teaching of corresponding concepts; 2) put an emphasis on the acquisition of lexical relations, paradigmatic (be a synonym/an antonym/a conversive of...) and syntagmatic (be an intensifier/a light verb of...), underlying paraphrastic links; 3) develop the necessary pedagogical tools—learner-friendly formalisms for encoding lexical relations and paraphrasing rules. This may seem almost obvious, but the actual practice is lagging far behind. One can only hope that the situation will change soon enough, with the paraphrase becoming a standard ingredient of language learning and teaching.

Acknowledgements

Many thanks to Stephanie Doyle Lerat, Lidija Iordanskaja and Igor Mel'čuk for their insightful remaks on a prefinal version of this paper.

Bibliography

Fuchs, C. 1980. *Paraphrase et théorie du langage. Contribition à une histoire des théories linguistiques contemporaines et à la construction d'une théorie énonciative de la paraphraase*. Thèse de doctorat. Paris: Université Paris VII.

Kahane, S. 2003. The Meaning-Text Theory. In: Agel, V. *et al*., eds, *Dependency and Valency. An International Handbook of Contemporary Research*, vol. 1. Belin/New York: De Gruyter; 546-570.

Keck, C. 2006. The Use of Paraphrase in Summary Writing: A Comparison of L1 and L2 Writers. *Journal of Second Language Writing* 15: 261-278.

Martinot, C. 2003. Pour une linguistique de l'acquisition. La reformulation: du concept descriptif au concept explicatif. *Langage et Société,* 2(104); 147-151.

Martinot, C. In press. Reformulations paraphrastiques et stades d'acquisition en français langue maternelle. *Cahiers de praxématique* 52: 1-18.

McInnis, L. 2009. *Analyzing English L1 and L2 Paraphrasing Strategies through Concurrent Verbal Report and Stimulated Recall Protocols*. MA Thesis. Toronto: University of Toronto.

Mel'čuk, I. 1992. Paraphrase et lexique: la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire. In: Mel'čuk, I. *et al*., *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*. Montréal: Presses de l'Université de Montréal; 9-59.

Mel'čuk, I. 1997. *Vers une linguistique Sens-Texte. Leçon inaugurale.* Paris: Collège de France.

Milićević, J. 2007. *La paraphrase. Modélisation de la paraphrase langagière*. Bern: Peter Lang.

Milićević, J. 2008. Paraphrase as a Tool for Achieving Lexical Competence in L2. In: Van Dalee, S. *et al*., eds, *Proceedings of the Symposium 'Complexity, Accuracy and Fluency in Second Language Use, Learing and Teaching'*. Brussels: Royal Belgium Academy of Science and Arts; 153-167.

Milićević, J. 2009. Lexical Functions and Paraphrasing Rules as a Link between L1 and L2. In: *Online Proceedings of the Conference 'First and Second Languages: Exploring the Relatinship in Pedagogy-related Contexts*. <http://www. education. ox, ac.uk/research/ resgroup/ alsla.aconf.php>

Polguère, A. 2004. La paraphrase comme outil pédagogique de modélisation des liens lexicaux. In : Calaque, E. & David, J., eds, *Didactique du lexique : contextes, démarches, supports*. Bruxelles: De Boeck; 115-125.

Russo, M. & Pippa, S. 2004. Aptitude to Interpreting : Preliminary Results of a Testing Methodology Based on Paraphrase. *META* 49/2; 409-432.

Thomas, A. 2008. La mesure des progrès lexicaux en FL2 avancé. In: Habert, D. & Laks, B eds, *Actes du Congrès Mondial de Linguistique Française 2008.* Paris: Institut de linguistique française; 587-597.

Wanner, L., ed. (1996). *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: Benjamins.

Žolkovskij, A. & Mel'čuk, I. 1967. O semantičeskom sinteze. *Problemy kiberetiki* 19: 177-238. [French translation: Sur la synthèse sémantique (1970). *TA Informations* 2; 1-85.]

Coreference of Deletions – The Case of Control

Giang Linh Nguy, Zdeněk Žabokrtský

Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800
{linh, zabokrtsky}@ufal.mff.cuni.cz

Abstract

In the present paper we focus on control as a subtype of anaphora. We work with the theory of control present within the dependency-based framework of Functional Generative Description (FGD[1]), in which control is defined as a relation of a referential dependency between a controller (antecedent – semantic argument of the main clause) and a controllee (anaphor – empty subject of the nonfinite complement (controlled clause)). First this paper presents the rule-based reconstruction of controllees, then, it discusses the perceptron-based determination of the controllees' antecedent. We evaluated our approach on data from the Prague Dependency Treebank 2.0, however, the rules and features of our system are supposed to be language independent and can be tested on other languages in the future.

Keywords

Control, control verb, controller, controllee, antecedent, anaphora, coreference.

1 Introduction

Anaphora resolution is widely studied for its important role in machine translation (MT). We believe that control as a subtype of anaphora can be helpful in MT as well. Consider following English sentences and their translations into Czech:

(1) John$_i$ told Mary$_j$ [Ø$_j$ to come].
Jan$_i$ řekl Marii$_j$, aby (ona$_j$) přišla.
Lit. John told Mary, so that (she) came.

[1] FGD bears numerous resemblances with Meaning-Text Theory, as discussed in (Žabokrtsky,2005)

(2) Mary$_i$ did not agree [Ø$_i$ to come].
Marie$_i$ nesouhlasila, že (ona$_i$) přijde.
Lit. Mary didn't agree, that (she) comes.

(3) Mary$_i$ hates John$_j$ [Ø$_j$ smoking].
Marie$_i$ nesnáší, když Jan$_j$ kouří.
Lit. Mary hates, when John smokes.

The mentioned examples show that the controlled clause can be expressed in one language by an infinitive verb or a gerund verb, whereas in another language, it can be expressed only by a finite verb.

Our work is divided into two steps: first we try to reconstruct controllees automatically by rules. After that we apply a perceptron-based ranker to identify the controllees' antecedent.

Before presenting our approach in Section 5, we discuss the theoretical background of control in Section 2, describe its annotation in the Prague Dependency Treebank in Section 3 and related works in Section 4. Section 6 summarizes experimental results. Conclusions and final remarks follow in Section 7.

To our knowledge, the present paper is one of the few papers, which deals with automatic coreference resolution of deletions in the case of control.

2 Theoretical Background

The terms used in our paper: *verb of control (control verb, governing verb)*, *controller (C-er)*, *controllee (C-ee)*[2], are known from Chomsky's framework of Government and Binding (Chomsky, 1981). In this paper, we work with Panevová's conception of Czech control (Panevová, 2000), in which control is understood in a broader way.

Panevová divides control into two groups: infinitive and nominalized constructions. The infinitive group is further divided into subgroups according to the position of the infinitive and the argument type of the controller. The nominalized group consists of only subgroups according to the argument type of the controller with the nominalized verb in the position of Patient and lacks the division according to the subject position.

Panevová et al. (2002) also presents another classification of control constructions: a combination of control verb and dependent verb both of which can be nominalized. An example of a control construction that can be expressed in all mentioned categories is: 1. *slíbit napsat dopis* (*to promise to write a letter),* 2. *slib napsat dopis* (*a promise to write a letter),* 3. *slíbit napsání dopisu* (*to promise writing of a letter),* 4. *slib napsání dopisu* (*a promise of writing of a letter).*

[2] In Ex. 1, the control verb is *told*, the controller is *Mary$_j$*, and the controllee is the covert argument *Ø$_j$* .

3 Control in the PDT 2.0

The Prague Dependency Treebank 2.0 (PDT 2.0) is a collection of Czech newspaper texts from 1990s to which a morphological annotation and annotation at two syntactic layers was assigned, at the so called analytical layer (a-layer, at which the surface shape of a sentence is reflected) and at the tectogrammatical layer (t-layer, which captures the linguistic meaning of the sentence); see (Hajič, 2006). Annotation of all three types is available for more than 49,000 sentences, consisting of more than 830,000 tokens.

At the t-layer, the meaning of the sentence is represented as a dependency tree structure (t-tree). Nodes of the tectogrammatical tree (t-nodes) are labeled with t-lemmas and dependency relations (functors[3], semantic roles) and enriched with valency annotation, annotation of semantically relevant grammatical meanings (grammatemes), annotation of topic-focus articulation, and annotation of coreferential relations including control constructions. The rules of t-layer annotation are described in (Mikulová et al., 2007).

In the PDT 2.0, two types of coreferential relations are distinguished: grammatical coreference, which is governed by rules of grammar of the given language, and textual coreference (Panevová, 1991). One kind of grammatical coreference is the relation between the controllee and the controller.

The annotation of control constructions, as of a special subtype of coreferential relations, in PDT 2.0 is based on Panevová's conception of control (Panevová, 2000). The up-to-date technical implementation of coreference in the PDT can be found in (Mikulová et al., 2007)[4]. Each t-node has the following coreferential attributes:

- `coref_gram.rf` – the ID of the grammatical antecedent (antecedents – in the case of conjunction)
- `coref_text.rf` – the ID of the textual antecedent(s)
- `coref_special = segm` (the antecedent is a segment of text) | `exoph` (the case of exophoric reference)

In (Kučová et al., 2003) and (Mikulová et al., 2007), the control classification given in Table 2 was extended by a new type of control – **quasi-control**. Quasi-control can be found within a complex (multi-word) predicate (Cinková, S. & V. Kolářová, 2004), where its verbal part and nominal part share some of their valency modifications. This sharing is called quasi-control.

(4) $Jan_{i\text{-}ACT}$ poskytl $Marii_{j\text{-}ADDR}$ [$Ø_{i\text{-}ACT}$ ochranu $Ø_{j\text{-}PAT}$].
Lit. John provided Mary protection.
$John_{i\text{-}ACT}$ provided [$Ø_{i\text{-}ACT}$ protection $Ø_{j\text{-}PAT}$] for $Mary_{j\text{-}ADDR}$.

[3] Functors represent the semantic values of syntactic dependency relations.

[4] (Panevová et al., 2002) and (Kučová et al., 2003) do not describe the current coreference implementation.

In Ex. 4, *'to provide protection'* is a complex predicate formed by a semantically empty verb *'to provide'* and a noun carrying the main lexical meaning of the entire phrase *'protection'*[5]. The omitted argument Actor of the noun *'protection'* refers to the verb's Actor *'John'* and the noun's non-expressed Patient refers to the verb's Addressee *'Mary'*.

At the t-layer, nodes correspond to autosemantic words only (including pronouns and numerals), prepositions and other functional words have no node in the tree. Besides t-nodes corresponding to surface tokens, in the tree there are newly established (reconstructed) t-nodes that have no counterpart in the outer shape of the sentence. These t-nodes have artificial t-lemmas prefixed by '#' (see their example in Fig. 1).

4 Related Work

There are many types of anaphora which have been actively studied in recent years, e.g. nominal and pronominal anaphora (Yang, 2008; Charniak & Elsner, 2009; Denis & Baldridge, 2009), bridging (indirect) anaphora (Poesio et al., 2004; Vieira et al., 2009), and zero anaphora (Kong & Zhou, 2010; Iisa & Poesio, 2011). Control as a subtype of zero anaphora was resolved in (Kučová et al., 2003) and (Ngụy, 2006).

Kučová et al. (2003) provided a rule set for some of control types: if the parent of an infinitive is a verb, then it is a control verb and the controllee refers to one of the control verb's arguments according to the list of control verbs. The list of control verbs was taken from the valency lexicon of Czech verbs VALLEX 1.0 and it includes only three types of control verbs: control verbs with Actor / Addressee / Patient controller[6]. The reported success rate of the rules was the following: ControlRuleACT 69.93%; ControlRuleADDR 88.64% and ControlRulePAT 33.33%.

Ngụy (2006) implemented a machine learning approach for the control coreference resolution, but the features given for training a decision tree were gained mainly from the list of control verbs extended by nominalized verbs. First a list of antecedent candidates was created. The list includes effective children of the controllee's effective grandparent[7] (except the controllee's effective parent); in cases of constructions 'to be resolved / able to do' effective children of controllees's great-grandparent; in cases of constructions 'It's possible / necessary to do' effective children of nodes with t-lemma 'možný / nutný / třeba'. Then, features of candidates were extracted. The features set was small, containing the candidate's t-lemma, functor, only one possible candidate option and the agreement of the candidate's and controllee's anaphor with the grandparent's category. Grandparent's categories are lists of control verbs and deverbal nouns. In addition to them are also ambiguous control verb lists – verbs with controllers of different functors or controller's and controllee's functors differ. The agreement of the candidate's and controllee's anaphor with the grandparent's category is then

[5] Its synonymous one-word predicate is 'to protect'.

[6] E.g. 'doporučit'-ADDR – to urge someone$_{i\text{-ADDR}}$ [Ø$_i$ to do something]; 'snažit se'-ACT – someone$_{i\text{-ACT}}$ to try [Ø$_i$ to do something]; 'poslat'-PAT – to send someone$_{i\text{-PAT}}$ [Ø$_i$ to do something]

[7] The "true governor" in terms of dependency relations.

detected by 18 rules. Using the described features, Ngụy trained a decision tree to decide whether a pair of controllee and antecedent candidate are coreferential. The success rate of her approach is 91.53%.

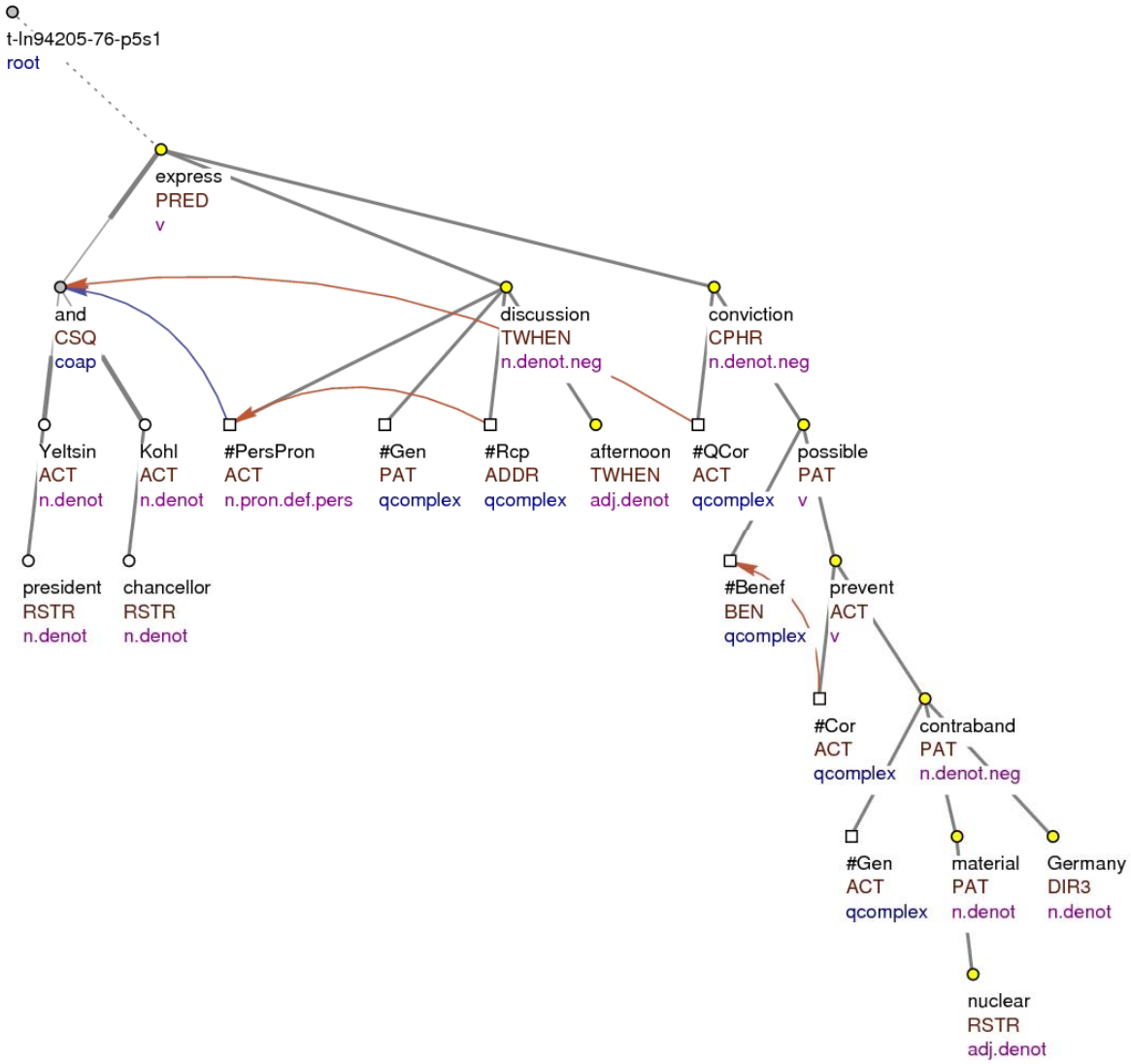


Figure 1: Simplified translated t-tree representing the sentence[8] *Prezident Jelcin a kancléř Kohl vyjádřili po odpoledních jednáních přesvědčení, že lze zabránit pašování jaderného materiálu do Německa*. (Lit.: *President Yeltsin and chancellor Kohl expressed after afternoon discussions the conviction, that it is possible to prevent from contraband of nuclear material to Germany*.)

[8] `#(Q)Cor` represents the (quasi-)controllee in control constructions; `#Benef` represents the beneficiary in control constructions; `#PersPron` represents overt and unstated personal or possessive pronouns (incl. the reflexives); `#Gen` represents a general participant absent at the surface level; and `#Rcp` stands for the omitted argument participating in a reciprocal relation.

5 Control Resolution

Our control coreference resolution task consists of two subtasks: first we have to identify anaphors, in our case the controllees; after that the antecedents, in our case the controllers have to be detected. The resolution for the first subtask is based on a list of t-lemmas of the controllees' effective parent. The second subtask is resolved by using a perceptron-based ranker inspired by Collins (2002).

5.1 Controllee Identification

The controllee identification process relies on the creation of a list of dependent verbs (deverbal nouns) for controllees and quasi-controllees from the training data. The list contains pairs of a dependent verb (noun) lemma and a controllee's functor. There are two independent procedures for identifying controllees and quasi-controllees. The procedure for controllees works as follows: for each infinitive, reconstruct a controllee with the functor, which either was found from the extracted list by the infinitive's lemma or was filled with 'ACT'.

In the case of quasi-controllees, the following simple rule was used: for each node with the functor 'CPHR'[9] and a lemma from the extracted list, reconstruct one or more quasi-controllees with different functors according to the list. [10]

5.2 Controller Detection

For the controller detection we use a simple scoring function: the optimal weight vector of which is estimated by averaged perceptron learning modified for ranking (Ngụy et al., 2009). The ranker is trained on the basis of feature vectors for a controllee and its possible antecedents. For every controllee a set of feature vectors containing only one positive instance and negative instances is formed. The positive instance includes features obtained from the controllee and its controller, whereas the negative ones are from the controllee and the non-coreferent phrase.

We consider three possible positions of the controller with respect to the controllee (Fig. 2):

1. the controller is the controllee's 'uncle' (the most frequent case)
2. the controller is the controllee's 'cousin' (in cases of control constructions 'It's possible / necessary to do')
3. the controller is a sibling of the controllees' effective grandparent (in cases of complex control construction[11])

[9] 'CPHR' is filled for the nominal part of a complex predicate.

[10] See the Example (4), in which two quasi-controllees occur: one with 'ACT' and another with 'PAT'.

[11] A complex control construction is meant as a construction of a complex control verb (predicate) + a dependent verb

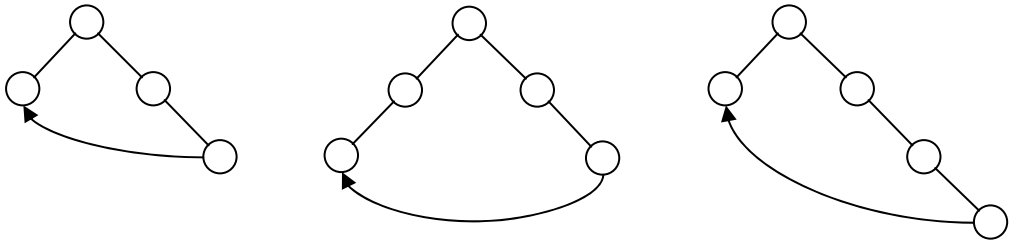


Figure 2: The tree representation of controllers' positions

The training features can be unary and related either to the controllee or to the candidate for the controller or to the controllee's effective parents (control verb and dependent verb), or they can be concatenated to represent the more complex relations between the controllee, the controller, the (complex) control verb (noun) and the dependent verb (noun). Altogether 30 features are used:

- Candidate (i): t-lemma, functor, tree position according to the controllee, semantic POS[12] (sempos), candidate's effective parent (ipar)'s t-lemma
- Controllee (j): t-lemma, functor
- Controllee's effective parent (jpar): t-lemma (lemma), functor (fun), sempos
- Controllee's effective grandparent (jpar2): t-lemma, functor, sempos
- Controllee's effective great-grandparent (jpar3): t-lemma, functor, sempos
- Concatenate(ipar_lemma, i_lemma): concatenation of the t-lemma of the candidate's effective parent and the candidate's t-lemma
- Concatenate(ipar_lemma, i_fun), Concatenate(jpar_lemma, i_fun, j_fun)
- Concatenate(jpar2_lemma, ipar_lemma), Concatenate(jpar2_lemma, i_fun), Concatenate(jpar2_sempos, i_fun), Concatenate(jpar2_lemma, i_fun, j_fun)
- Concatenate(jpar2_lemma, jpar_lemma, i_fun, j_fun), Concatenate(jpar2_lemma, jpar_sempos, i_fun), Concatenate(jpar2_lemma, jpar_sempos, i_fun, j_fun),
- Concatenate(jpar3_lemma, jpar2_lemma, i_fun, j_fun),
- Concatenate(jpar3_lemma, jpar2_lemma, jpar_lemma, i_fun, j_fun)
- Concatenate(jpar2_lemma, ipar_lemma, jpar_lemma, i_fun, j_fun)
- Concatenate(jpar2_lemma, ipar_lemma, i_lemma, jpar_lemma, j_lemma, i_fun, j_fun)

[12] Semantic parts of speech correspond to the basic onomasiological categories.

6 Evaluation

Manually annotated tectogrammatical trees from the PDT 2.0 are used both for training and evaluation purposes. We employ the standard division of the PDT 2.0 into three parts: 80% of data is used for training, 10% for development testing and 10% for evaluation testing. The training data contains 6,598 controllees, 874 development test data and 907 evaluation test data. In the evaluation we used standard metrics with precision, recall and f-measure (Table 1) for controllee identification and controller detection.

Precision = N_c / N_e	Recall = N_c / N_t	F-measure = 2 x P x R / (P + R)

N_c	Number of correctly identified controllees/controllers
N_e	Number of identified controllees/controllers
N_t	Number of all controllees/controllers

Table 1: Evaluation metrics for the control resolution

We applied the following baseline rule for controller detection: for each controllee, select its 'uncle' with functor 'ACT' as its controller. The scores of rules for the controllee and quasi-controllee identification and the baseline rule and ranker for controller detection are given in Table 2.

	P	R	F
Cor.Ident.Rule	83.381%	86.222%	84.778%
QCor.Ident.Rule	86.219%	85.915%	86.067%
Coref.Baseline	56.065%	57.351%	56.701%
Coref.Ranker	82.161%	84.046%	83.093%

Table 2: Results for the control resolution

The errors of controllee (Cor) identification arise in cases: dependent verb is nominalized (14.525%); Cor was not annotated; Cor was annotated with `#PersPron` or `#Gen` instead. The problem with quasi-controllee (QCor) identification was the recognition of its functor. If the correct recognition of QCor's functor is not in the task, then the f-measure is 96.075%.

The success rate of the automatic control coreference resolution depends on the previous subtask, the controllee identification. If the control coreference ranker is tested on golden trees (with manually annotated controllees), then it achieves the f-measure of 96.246% and outperforms the system of (Ngụy, 2006). The errors of the ranker occur when the controller is a verb or an adjective; or the controller is in another position than those given in Fig. 2.

7 Conclusions

In this paper we report two systems for the automatic resolution of control. One addresses controllee identification and uses hand-written rules based on lists of control verbs. The second system works with a ranker for controller detection and achieves the success rate of 83.093%. The result can approach nearly 96.246% if the controllee identification resolution is significantly improved.

Acknowledgements

The work on this project was supported by the grants MSM0021620838, GA405/09/0729, MŠMT ČR LC536 and GAUK 4383/2009.

Bibliography

Charniak, E. & M. Elsner. 2009. EM works for pronoun anaphora resolution. In: *Proceedings of EACL '09*, 148–156.

Chomsky, N. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Mouton de Gruyter.

Cinková, S. & V. Kolářová. 2004. Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In: *Korpusy a korpusová lingvistika v zahraničí a na Slovensku*.

Collins, M. 2002. Discriminative Training Methods or Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In: *Proceedings of EMNLP '02*, Vol. 10, 1–8.

Denis, P. & J. Baldridge. 2008. Specialized models and ranking for coreference resolution. In: *Proceedings of EMNLP '08*, 660–669.

Hajič J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský & M. Ševčíková-Razímová. 2006. *Prague Dependency Treebank 2.0*. Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, www.ldc.upenn.edu.

Iida, R. & M. Poesio. 2011. A Cross-Lingual ILP Solution to Zero Anaphora Resolution. In: *Proceedings of ACL '11*, 60–69.

Kong, F. & G. Zhou. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In: *Proceedings of EMNLP '10*, 882–891.

Kučová, L., V. Kolářová, Z. Žabokrtský, P. Pajas & O. Čulo. 2003. *Anotování koreference v Pražském závislostním korpusu*, Prague: MFF UK, TR-2003-19.

Mikulová, M., A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Ševčíková, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá

& Z. Žabokrtský. 2007. *Annotation on the tectogrammatical level in the Prague Dependency Treebank*. Technical report no. 2007/3.1, ÚFAL, Charles University.

Ngụy, G. L. 2006. *Proposal of a Set of Rules for Anaphora Resolution in Czech*. Master's thesis, Faculty of Mathematics and Physics, Charles University.

Ngụy, G. L., V. Novák & Z. Žabokrtský. 2009. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In: *Proceedings of the SIGDIAL 2009 Conference*, 276–285.

Panevová, J. 1991. Koreference gramatická nebo textová? In: *Etudes de linguistique romane et slave*. Krakow.

Panevová, J. 2000. More Remarks on Control. In: *Prague Linguistic Circle Papers*, Vol. 2, 101–120. Amsterdam/Philadelphia: Benjamins Publishing House.

Panevová, J., V. Řezníčková & Z. Urešová. 2002. The theory of control Applied to the Prague Dependency Treebank (PDT). In: *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks*, 175–180. Università di Venezia.

Poesio, M., R. Mehta, A. Maroudas & J. Hitzeman. 2004. Learning to resolve bridging references. In: *Proceedings of ACL '04*.

Sgall, P. 1967. *Generativní popis jazyka a česká deklinace*. Praha, Academia.

Sgall, P., E. Hajičová & E. Panevová. 1986. *The Meanings of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, D. Reidel Publishing Company - Praha, Academia.

Sgall, P. et al. 1986. *Úvod do syntaxe a sémantiky*. Praha, Academia.

Vieira, R., E. Bick, J. Coelho, V. Muller, S. Collovini, J. Souza & L. Rino. 2009. Semantic tagging for resolution of indirect anaphora. In: *Proceedings of SigDIAL '06*, 76–79.

Yang, X., J. Su & C. L. Tan. 2008. A twin-candidate model for learning-based anaphora resolution. In: *Computational Linguistics*, Vol. 34, 3, 327–356.

Žabokrtský, Z. 2005. Resemblances between Meaning-Text Theory and Functional Generative Description. In: *Proceedings of the 2nd International Conference of Meaning-Text Theory*, 549–557. Slavic Culture Languages Publishers House, Moskva, Russia, ISBN 5-9551-0094-6.

Is Linear Order Derived?

Timothy Osborne

tjo3ya@yahoo.com

Abstract

The paper casts a critical eye on those dependency grammars (DGs) that view linear order as derived from hierarchical order. It argues that MTT is such a DG insofar as language synthesis begins with a semantic representation, progresses through syntactic representations that lack linear order, and then proceeds to the morphological and phonological representations that include linear order. The difficulty that the intermediate syntactic representations generate is they necessitate that certain phenomena of syntax be subjected to analyses that ignore linear order. This necessity is problematic in various areas, such as idiosyncratic meaning and coordination. Monostratal DGs are not faced with these difficulties, since they do not posit intermediate syntactic strata that lack linear order.

Keywords: coordination, hierarchical order, idioms, linear order, monostratal, multistratal

1 Introduction

Is linear order derived? This question seems to be (at least tacitly) answered positively by many dependency grammars (DGs) insofar as many DGs view hierarchical order as being less deep than linear order. These DGs are backed to some extent by the following crucial passage from Tesnière:

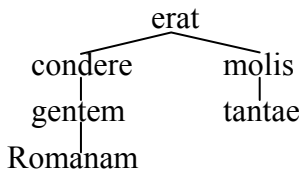
> De ce point de vue, nous pouvons dire, en reprenant notre definition du début […] pour la préciser et la developer, que parler une langue, c'est en transformer l'ordre structural en ordre linéaire, et inversement que comprendre une langue, c'est en transformer l'ordre linéaire en ordre structural.
> (Tesnière, 1959/69:19).

According this passage, a speaker transforms structural order (=hierarchical order) to linear order. Such transformation necessitates that the hierarchical order precede the linear order. But if the hierarchical order precedes the linear order, then the hierarchical order is primitive and the linear order is (at least in some sense) derived (for the speaker). The understanding of speech production implied by this view is that a more or less complete hierarchical structure of morphological units exists in cognition immediately before that speaker begins an utterance. Considering the extended length and major complexity of many utterances, one can

question whether this understanding of speech production is accurate. The burden placed on the cognition of a speaker (to first construct a more or less complete hierarchical structure in cognition before speech production begins) would be extreme. A more plausible understanding of speech production assumes that hierarchical order is established (in an abstract sense) online in tandem with linear order.

That many DGs indeed derive linear order from hierarchical order is implied by the widespread trees that encode hierarchical order alone (not linear order as well). Tesnière seems to have initiated this practice; the majority of tree structures in *Éléments de syntaxe structurale* convey hierarchical order only. The structure of the Latin sentence *Tantae molis erat Romanam condere gentem* 'It was such a massive task to establish the Roman race', for instance, is given as follows:

(1)

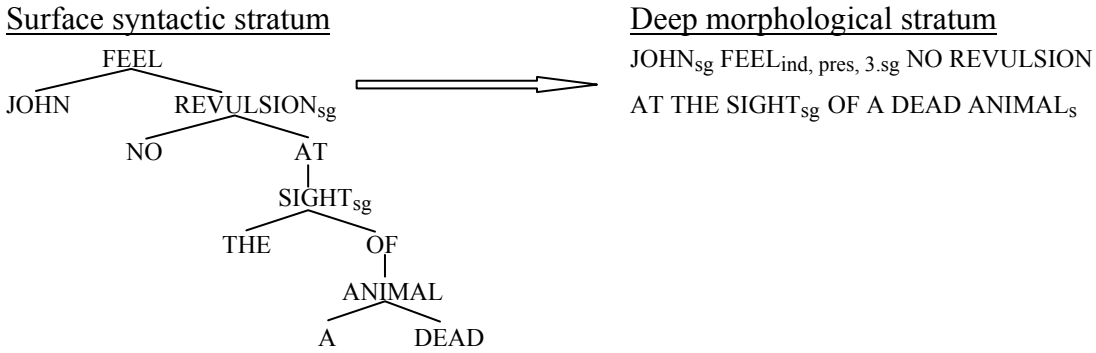


(Tesnière, 1959/69:20)

Tesnière's proposal can be interpreted to mean that the Latin speaker has this hierarchy of words present in cognition immediately before he/she begins to utter the corresponding sentence. Speech production entails the transformation of hierarchical structure to linear order.

Meaning to Text Theory (MTT: Mel'čuk, 1988, 2003; Kahane, 2003) can be interpreted as following this practice. MTT posits two strata of syntax, each of which encodes hierarchical order but not linear order. Viewed from the perspective of language synthesis, linear order appears first in the deep morphological stratum after the surface syntactic structure is mapped to it. The following representations have been adapted (syntactic functions omitted) from Kahane (2003:557f.):

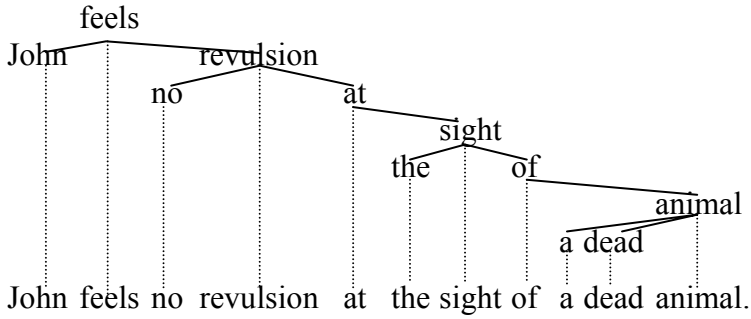
(2)



The surface syntactic stratum is a tree that contains all the lexemes of the actual utterance *John feels no revulsion at the sight of a dead animal*, but lacks linear order. Again from the perspective of synthesis, linear order is appearing first as the unordered tree is mapped to the deep morphological stratum, which now encodes linear order but lacks hierarchical order.

Monostratal DGs reject this view of how linear order is established (e.g. Hays, 1964; Hudson's Word Grammar, 1990; Starosta's Lexicase, 1988; Groß, 1999; Osborne et al., in press). These grammars assume a single syntactic stratum, whereby this stratum encodes both hierarchical and linear order simultaneously. Such systems might produce a tree like the following one for the sentence *John feels no revulsion at the sight of a dead animal*:

(3)



This tree encodes both hierarchical and linear order simultaneously. The underlying premise behind monostratal syntax in this area is that both ordering dimensions are always simultaneously present in syntax. The one dimension is not more basic than the other, which means that the one cannot be derived from the other. While monostratal systems may on occasion examine or reference the dominance dimension alone, doing so for them does not entail that they view such a representation as corresponding to anything that is real (and thus worthy of receiving an analysis at a putatively distinct level of syntactic representation).

This contribution considers some conceptual and empirical arguments that support monostratal syntax. In particular, some observations from idioms, coordinate structures, and the general trend associated with Construction Grammar (CxG) are brought to bear on the debate. The main objection raised by anonymous reviewers is addressed in the penultimate section.

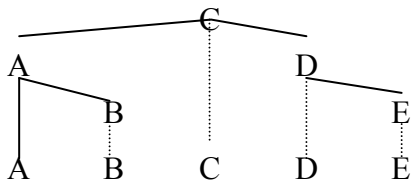
2 Construction Grammar

Insights coming from construction grammars in recent years have been influencing our understanding of syntax and grammar in significant ways (e.g. Fillmore et al., 1988; Kay & Fillmore, 1999; Goldberg, 2006). The strict compositionality of generative grammars (in the GB/MP tradition) is under debate, and the systems that derive syntactic structure incrementally are being challenged, whereby representations have come to play a greater role in our knowledge of how meaning is encoded and conveyed. The primary unit of construction grammars is of course the construction, which is understood to be a conventionalized means of connecting meaning to sound and sound to meaning. Constructions are stored units that consist of information from a variety of domains, i.e. logical-semantic, lexical, functional, prosodic, purely syntactic, pragmatic, etc. The key aspect of constructions in the current context is that many of them are widely acknowledged to include linear order (e.g. Lakoff, 1987:489; Kay & Fillmore, 1999:3; Croft, 2001:196f.). In other words, many constructions impose both linear and hierarchical order on their parts. Since constructions are stored units, i.e. they are stored on the lexicon-syntax continuum, and these stored units often necessitate a specific linear order to their parts, linear order is inseparable from constructions. The point, then, is that syntactic representations that encode hierarchical order alone deliver an inaccurate impression about the nature of these constructions. Tree structures that encode both hierarchical and linear order simultaneously, however, are not faced with this difficulty.

An important aspect of the DG view of syntactic (and morphosyntactic) structure bears on these matters. Holmes & Hudson (2005) argue that dependency-based structures are particularly compatible with the tenets of Construction Grammar. Indeed, adopting the DG *catena* (O'Grady, 1998; Osborne, 2005; Osborne et al. in press) as the fundamental unit of syntax and morphosyntax, one is in a position to give many constructions a concrete

expression in surface (morpho)syntax. This point will be illustrated here with respect to a couple of lexical and syntactic constructions that necessarily impose linear order on their parts. The catena is defined as A WORD OR A COMBINATION OF WORDS THAT IS CONTINUOUS WITH RESPECT TO DOMINANCE. This definition identifies any dependency tree or subtree of a dependency tree as a catena. The following abstract tree is used to illustrate the concept; the letters represent words:

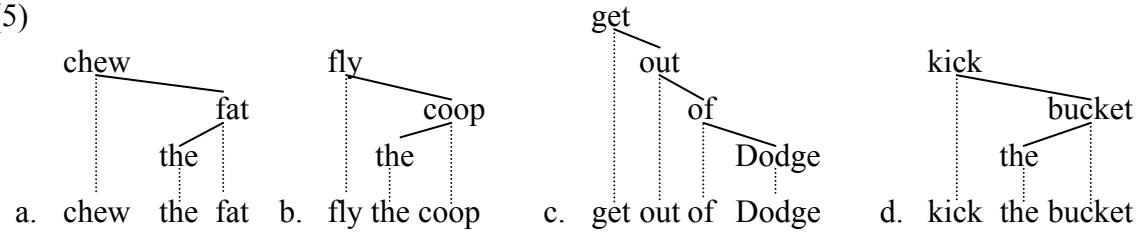
(4)



This structure contains 15 distinct catenae: A, B, C, D, E, AB, AC, CD, DE, ABC, ACD, CDE, ABCD, ACDE, and ABCDE. It also contains 16 distinct non-catenae: AD, AE, BC, BD, BE, CE, ABD, ABE, ACE, ADE, BCD, BCE, BDE, ABCE, ABDE, and BCDE. As the number of words increases, the percentage of non-catena combinations increases. The importance of the catena for constructions is that many units that arguably qualify as constructions are stored as catenae.

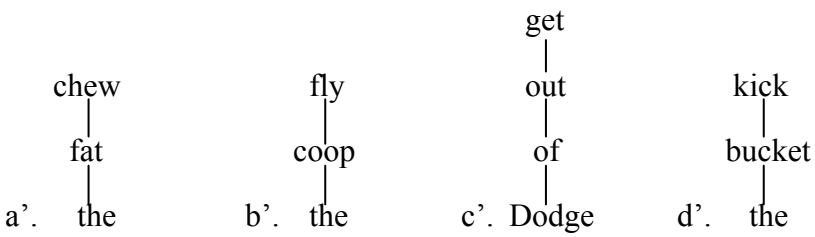
Construction grammars view idiosyncratic expressions of all sorts as lexically fixed constructions (e.g. Croft, 2001:15). One particular group of idioms is of particular interest to the current discussion, e.g. *chew the fat, fly the coop*, *get out of Dodge*, *kick the bucket*. The words of these idioms are syntactically fixed insofar as the order in which they appear is set. First, note that these idioms are catenae:

(5)



Unlike many "mobile idioms" (see Horn, 2003), these idioms cannot be altered in any major way (without losing their idiomatic meaning). For instance, passivization is not possible: **The fat was chewed by us*, **The coop was flown by the thief*, **Dodge was got out of by us*, **The bucket was kicked by him*. The noun (phrase) cannot be topicalized: **..and the fat we chewed*, **…and the coop the thief flew*, **…and Dodge we got out of*, **…and the bucket he kicked*. The nouns cannot be modified/questioned: **Which fat did they chew?*, **Which coop did the thief fly?*, **Which bucket did he kick?* The importance of these data for the discussion should be apparent. Linear order is clearly part of these idiom catenae. Unordered dependency trees deliver an inaccurate impression about the nature of these idioms, e.g.

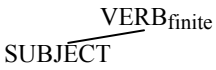
(5)



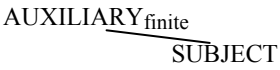
Linear order is absent from these hierarchies. Models of syntax (like MTT) that posit one or more syntactic strata that include hierarchical order but exclude linear order in this manner are delivering an inaccurate impression about the nature of these idiom constructions. They are inaccurately suggesting that the linear order of the words of these idioms is flexible. This situation should motivate one to question the legitimacy of the intermediate, syntactic strata.

Major syntactic constructions that are syntactically fixed but lexically free also encode linear order. For instance the SV construction (e.g. Kay & Fillmore, 1999:12; Sag 2010) and aux-inversion construction (Kay & Fillmore, 1999:18; Goldberg, 2006:Ch. 8; Goldberg, 2009:110ff.; Sag 2010:46f.) of English:

(6) SV Construction



(7) Aux-inv. construciton

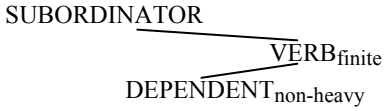


The SV construction shown in (6) is perhaps the most frequently occurring syntactic construction of English. Certainly the majority of clauses in this paper contain this construction. The SV construction is associated with neutral declarative illocutionary force. In the current context, the key point about the SV construction is that the subject must precede the finite verb, which means linear order is an inseparable part of the construction. The same insight is applicable to the aux-inversion construction. This construction is associated with interrogative or affective illocutionary force. Goldberg (2009:112) posits that the main semantic content expressed by the aux-inversion construction is "non-positivity".

The major syntactic constructions of German most responsible for establishing word order in clauses are also constructions that encode linear order. The catenae of the V2 and VF constructions are represented schematically as follows:

(8) V2 construction

(9) VF construction



The V2 construction requires that one and only one pre-dependent attach to the finite verb. As long as this pre-dependent has some semantic content, its syntactic status is unconstrained; it can be an argument, an adjunct, or part of the predicate. The V2 construction is associated with matrix clauses that have neutral illocutionary force. The VF construction also encodes linear order; the finite verb daughter of a subordinator must be a post-dependent. The key trait of the VF construction is that it lacks independent illocutionary force (since it is the mark of an embedded clause). Another important aspect of the VF construction is that it requires non-heavy dependents of the finite verb to appear as pre-dependents, whereby only relatively heavy dependents can appear as post-dependents (e.g. clauses, heavy *zu*-phrases, heavy PPs).

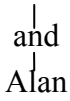
The Construction Grammar understanding of the constructions schematized in (6-9) is that the linear order shown is an inseparable part of each construction. Models of syntax like MTT that acknowledge no level of representation that simultaneously encodes both hierarchical and linear order are therefore challenged. These models are, namely, faced with the difficult question of discerning the appropriate level of representation where these constructions exist and can be represented. Intermediate hierarchies that only encode hierarchical order ignore a primary trait of these constructions, namely that they cannot exist without linear order.

3 Coordination

The message delivered by the previous section is that many constructions necessarily encode linear order and that tree structures omitting linear order deliver an inaccurate impression of the nature of syntax and the lexicon. This section examines the issue from another perspective, that of coordination. Assuming the existence of intermediate levels of syntax lacking linear order forces one to subject coordinate structures to an implausible hierarchical analysis. This analysis is inconsistent with various empirical aspects of coordination.

Multistratal MTT grants coordinate structures a hierarchical organization that is essentially the same as that of non-coordinate structures. The following (b)-example is from Mel'čuk (2003:217):

(10) a. Leo and Alan

b.


c. [[Leo] and [Alan]]

According to Mel'čuk, the dependency structure of (10a) is as shown in (10b). Coordinate structures therefore receive an analysis that is close to the analysis of non-coordinate structures. The important thing to acknowledge in this regard is that the analysis shown in (10b) is forced by theory-internal necessity of the MTT multistratal system. MTT's surface syntactic stratum lacks linear order at the same time that it includes a complete hierarchy of lexical items. In order for the surface syntactic stratum to accommodate coordinate structures, they must receive an analysis that allows them to be integrated into the unordered tree. An analysis like that shown in (10b) can accomplish this task.

The problem with the analysis in (10b), though, is that it is contradicted by a number of empirical facts. The analysis in (10c) is more plausible, since it assumes that coordinate structures are organized in terms of constituency (not in terms of dependency). Before examining two empirical arguments that support (10c) over (10b), consider how Mel'čuk (2003:217) motivates (10b). He points to facts from Bantu languages and from Nias, an Indonesian language, whereby the first conjunct of a coordinate structure often receives a morphological marking that the other conjuncts do not receive. Based on this observation, Mel'čuk concludes that the first conjunct must be the root of the coordinate structure, the other conjuncts being subordinate to it. While this reasoning is understandable given the multistratal nature of MTT, it is not necessary or plausible if one assumes a monostratal system, since a monostratal system puts linear order on equal footing with hierarchical order and can hence address the asymmetries that Mel'čuk observes in terms of linear ranking. The first conjunct is more prominent than the following conjuncts by virtue of the simple fact that it precedes the other conjuncts. In other words, the monostratal system grants linear order a similar power of organization as it does hierarchical order. This possibility is not available for MTT, since the MTT understanding of language synthesis views linear order as derived.

A DG analysis of coordinate structures like the one shown in (10c) is preferable. This analysis takes coordinate structures to be fundamentally different from non-coordinate structures. While non-coordinate structures are organized with respect to both dimensions simultaneously, coordinate structures are organized along the horizontal dimension only. They therefore behave in a manner that is distinct from non-coordinate structures. As stated, a

number of empirical facts suggest that (10c) is correct. One of these is the nesting that occurs in coordinate structures. These nestings are organized in terms of constituency, e.g.

(11) a. [Sam] and [Fred] and [Susan] arrived late.
b. [[Sam] and [Fred]] and [Susan] arrived late.
c. [Sam] and [[Fred] and [Susan]] arrived late.

The groupings indicated by the brackets carry meaning depending on who arrived with whom, but these meanings cannot be accommodated with pure dependency structures of the sort shown in (10b). Instead, one can assume constituency as the underlying principle of organization, and in order to acknowledge constituency in this manner, one has to assume that linear order is basic.

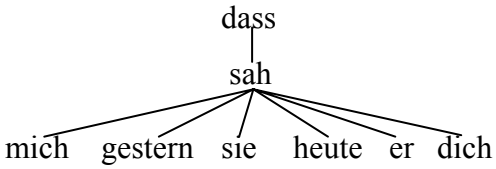
A similar type of datum involving coordination occurs when the conjuncts fail to qualify as complete subtrees (=constituents), as the following clause from German illustrates:

(12) dass [er mich heute] und [sie dich gestern] sah
that he me today and she you yesterday saw
'that he saw me today and she you yesterday'

The structure of the conjuncts in this case is flat. In order to accommodate this flatness, an analysis along the lines of (10b) might assume a deletion mechanism that reduces the conjunct(s) down to their surface form. Such a mechanism would, however, be contrary to the non-derivational nature of most DGs.

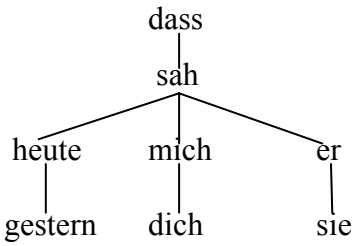
Or if one nevertheless chooses to subject (12) to a hierarchical analysis that lacks linear order, one might end with something like (12'):

(12')



The difficulty now is that it is not at all clear how this representation can be mapped to a representation that shows linear order. An alternative analysis might try the following hierarchy:

(12'')



While this hierarchy correctly groups the words bearing the same syntactic function together, it is still faced with the difficulty of linearizing the words of the conjuncts in a plausible manner. Note that if the hierarchy of words were to be maintained and shown in a tree that also shows linear order, projectivity violations would be present on a grand scale.

Monostratal DGs are challenged to a lesser extent by these issues, since they encode hierarchical and linear order simultaneously. They can assume that the conjuncts of coordinate structures are organized in terms of constituency along the linear dimension. See Osborne 2006a, 2006b, 2008.

4 The objection

Two anonymous reviewers have objected that the message of this paper misunderstands and misrepresents the nature of the MTT model. They emphasize that the model is "bi-directional". The blanket statement that MTT views linear order as derived is wrong, since the model operates in both directions depending on whether it is interpreted from the synthesis or analysis perspective. From the synthesis perspective (= the speaker's perspective) the model begins with SemR and progresses in sequence through DSynR, SSynR, DMorphR, SMorphR, and DPhonR to SPhonR. But from the analysis perspective (= the listener's perspective), the model begins with the SPhonR and progresses in sequence through DPhonR, SMorphR, DMorphR, SSynR, and DSynR to SemR. Thus the claim that hierarchical order is more basic than linear order is simply wrong for the analysis perspective.

None of the claims or points raised and discussed above misrepresent the bi-directionality of the MTT model. The discussion surrounding the quotation from Tesnière in the introduction makes it clear that the message is concerned primarily with the synthesis perspective, not so much with the analysis perspective. That is, the synthesis perspective of MTT, which sees linear order being derived from hierarchical order, is implausible. A model that necessitates syntactic representations that lack linear order misrepresents the nature of many phenomena. A solid understanding is hampered by the necessity to force representations on the syntactic strata that lack an integral aspect of the phenomena under scrutiny, namely linear order. The analysis of coordination is perhaps the most vivid example of this problem, as discussed above. In order to produce a structural analysis of coordinate structures on the syntactic strata (which, again, lack linear order), one has to impose a dependency-based analysis onto the conjuncts. There is, however, considerable empirical evidence suggesting that constituency is the principle that organizes the conjuncts of coordinate structures. In order to acknowledge the role of constituency, however, one has to acknowledge linear order.

While the discussion above has focused primarily on the synthesis perspective of the MTT model, the basic message is also valid for the analysis perspective. The current stance is that the analysis perspective of MTT is implausible insofar as it takes hierarchical order to be derived from linear order. The same sorts of difficulties are going to arise in this area. The necessity to produce analyses of certain phenomena while neglecting the role of hierarchical organization is going to deliver an inaccurate impression of the nature of these phenomena.

5 Concluding remarks

Some remarks about MTT and related DGs conclude this contribution. According to Kahane (2003:546), MTT was developed in the 1960s in Moscow as the theoretical background for a project to computationally parse Russian. The fact that Russian was the original object of inquiry is important, since as a Slavic language, the word order of Russian is of course quite free compared to that of other languages like English and German. This aspect of Russian likely influenced the development of the theory. In other words, the relatively free word order of Russian found an expression in MTT insofar as linear order could more easily be granted

secondary status. In this vein, it is worth taking note of Mel'čuk's (1988:4) comments concerning the creation of phrase structure (=PS) grammars of the Chomskian tradition:

> Even though it sounds a bit too Whorfian, I am fairly sure that PS-syntax could not have been invented and developed by a native speaker of Latin or Russian…To promote PS-syntax, one has to be under the overall influence of English, with its rigid word order and almost total lack of syntactically driven morphology. (Mel'čuk 1988:4)

While I agree with this statement, its expression raises a flag concerning the influences surrounding the creation of MTT. One should inquire, namely, whether a theory like MTT was originally conceivable because the object of inquiry was Russian (with its relatively free word order), and not, say, English (with its relatively fixed word order).

The message delivered with this contribution is that in many cases, linear order is not derived, but rather it is primitive. It is often encoded as an integral part of a construction. The relatively free word order of a language like Russian is explained in part by the assumption that relatively few constructions in Russian encode linear order, whereas relatively many constructions in a language like English or German do encode linear order. Russian, for instance, lacks an SV construction, a V2 construction, and a VF construction. Lacking these constructions, word order in Russian relies more on functional principles associated with information structure.

A final comment speculates about the inability of DG to establish a foothold in many linguistics circles. Compared to constituency-based syntax, dependency-based syntax is by most objective measures truly minimal. The expectation should therefore be that dependency-based syntax is preferred. The fact that it clearly is not – constituency-based systems still occupying center stage in theoretical linguistics (at least in North America) – should motivate one to question what DGs have overlooked. The unordered trees frequently produced by many DGs may play a negative role in this regard. Established and aspiring linguists who are first exposed to dependency-based theories encounter these unordered trees. The impression one gets is that DG is vague, the tree structures lacking linear order. Constituency-based systems, in contrast, necessarily encode linear order and thus linear order has been front and center for them from the start. DGs can perhaps learn a lesson in this area. Tree structures that encode linear order may be a more effective tool for conveying the advantages of dependency-based syntax, and given the message of this contribution, the ordered tree structures now enjoy greater theoretical legitimacy.

Bibliography

Croft, W. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. New York: Oxford University Press.

Fillmore C., P. Kay & M. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64:501-538.

Goldberg, A. 2006. *Constructions at work*. Oxford: Oxford University Press.

Goldberg, A. 2009. The nature of generalizations in language. *Cognitive Linguistics* 20:93-127.

Groß, T. 1999. *Theoretical foundations of dependency syntax*. Munich: Iudicium.

Hays, D. 1964. Dependency theory: A formalism and some observations. *Language* 40:511-525.

Holmes, J. & R. Hudson. 2005. Constructions in Word Grammar. J-O. Ostman & M. Fried (eds.), *Construction grammars: Cognitive grounding and theoretical* extensions, 243-272. Amsterdam: Benjamins.

Horn, G. 2003. Idioms, metaphors, and syntactic mobility. *Journal of Linguistics* 39.245-273.

Hudson, R. 1990. *An English Word Grammar*. Oxford: Basil Blackwell.

Kahane, S. 2003. The Meaning-Text Theory. V. Ágel et al. (eds.), *Dependency and valency: An international handbook of contemporary research*, vol. 1, 546-569. Berlin: Walter de Gruyter.

Kay, P. & C. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. *Language* 75:1-33.

Lakoff, G. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: The University of Chicago Press.

Mel'čuk, I. 1988. *Dependency syntax: theory and practice*. Albany: State University of New York Press.

Mel'čuk, I. 2003. Levels of dependency description: Concepts and problems. V. Ágel et al. (eds.), *Dependency and valency: An international handbook of contemporary research*, vol. 1, 188-229. Berlin: Walter de Gruyter.

O'Grady, W. 1998. The syntax of idioms. *Natural Language and Linguistic Theory*16:79-312.

Osborne, T. 2005. Beyond the constituent: A dependency grammar analysis of chains. *Folia Linguistica* 39:251-297.

Osborne, T. 2006a. Shared material and grammar: A dependency grammar theory of non-gapping coordination. *Zeitschrift für Sprachwissenschaft* 25:39-93.

Osborne, T. 2006b. Parallel conjuncts. *Studia Linguistica* 60(1):64-96.

Osborne, T. 2008. Major constituents: And two dependency grammar constraints on sharing in coordination. *Linguistics* 46(6):1109-1165.

Osborne, T., M. Putnam & T. Groß (in press). Catenae: Introducing a novel unit of syntactic analysis. *Syntax*.

Sag, I. 2010. Sign-Based Construction Grammar: An informal synopsis. H. Boas & I. Sag (eds.), *Sign-Based Construction Grammar*, 39-160. CSLI.

Starosta, S. 1988. *The Case for Lexicase: An Outline of Lexicase Grammatical Theory*. New York: Pinter Publishers.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Tesnière, L. 1969. *Éléments de syntaxe structurale*, 2nd edition. Paris: Klincksieck.

Meanings and Ontological Categories of the Russian Word vpechatlenie 'impression'

Elena Paducheva

Russian Academy of Sciences, VINITI,
elena.paducheva@yandex.ru

Abstract

The word *впечатление* 'impression' in modern Russian is morphologically non motivated; it is characterized by unique combinability and non regular polysemy. V.V.Vinogradov (Vinogradov, 1994) cites Lev Tolstoy, who had chosen the word *впечатление* in order to illustrate his idea that it is impossible to describe the meaning of a word resorting to whatever other words. The paper aims at analyzing the word *впечатление* in the framework of systemic lexicology.

Keywords

Thematic class, ontological category, diathesis, combinability, word formation pattern

1 Puzzles connected with the word *впечатление*

Semantically, the word *впечатление* should belong to the thematic class of emotions: "something made an impression upon me" means that I am now in a special emotional (or mental – in any case, psychological) state. In fact, the word *впечатление* is often used in the context of other nouns of emotion (this and many other examples below are taken from the National Corpus of Russian – http://www.ruscorpora.ru):

> Впечатления меняются одно за другим: недоумение, удивление и, наконец, восхищение лихой напористостью авангардных художников. [«Вокруг света», 2004.07.15]

However, as to its linguistic behavior, the word *впечатление* is different from typical emotion nouns.

Puzzle 1. **Word formation pattern**

According to its morphological structure, *впечатление* is a verbal noun. The only verb in the dictionaries of modern Russian with which it can be morphologically correlated is the verb *впечатлить* 'to impress'. But *впечатлить* is defined as *произвести впечатление* 'to make an impression' and the word combination

произвести впечатление gives no clue to the semantic relationship between *впечатление* and *впечатлить*.

In fact, many nouns of emotion form regular proportions with causative verbs of emotion: *удивить = вызвать удивление*, *смутить = вызвать смущение*, *встревожить = вызвать тревогу*, etc. But the noun *впечатление* does not enter this list:

Книга впечатлила – *Книга *вызвала* впечатление.

One should say –

Книга *произвела* впечатление.

As far as in the 19th century the verb *производить* was widely used in the function of a "verbalizer" for nouns of emotion:

Эта речь *произвела удивление*, не менее предшествующей. [П. В. Анненков. Записки о французской революции 1848 года (1848)]

Ужели мой друг думал что чин Патриция во мне *произвел радость*. [А. Н. Радищев. [Положив непреоборимую преграду...] (1790)]

Он ожидал, что его заявление просто *произведет тревогу*, но оно не произвело ничего. [Н.С. Лесков. На ножах (1870)]

Later on, the verb *производить* is used more rarely and *вызывать* can always be substituted for *производить*:

Боже мой! — какой мы *произведем перепуг*! В искусстве нами любуются. [К. А. Федин. Первые радости (1943-1945)] [= *вызовем*];

Когда картина была выставлена, она *произвела смятение* среди лондонцев. [К.Г.Паустовский. Золотая роза (1955)] [= *вызвала*];

Они нас не ждут с этой стороны, наше появление *произведет панику*. [Н. С. Гумилев. Записки кавалериста (1914-1915)] [= *вызовет*].

However, in the context of *впечатление* it is impossible to substitute *вызвать* for *произвести*. Thus, *впечатление* is not related to causative *впечатлить* along the same model that correlates *удивление* with *удивить*.

Let us take the verb *впечатлиться* – a reflexive (or medial) counterpart of *впечатлить*. This verb is not present in MAC, but 14 examples were found in the National Corpus of Russian, all of them perfectly acceptable, such as:

Впечатлившись услышанным и увиденным, С. В. Степашин пообещал, что … . [«Встреча» (Дубна), 2003.04.23]

Но более всего ревнительские издания *впечатлились* результатами суда в г. Приозерске [«Церковный вестник», 2002.11.10]

Now, decausative verbs of emotion can be connected with corresponding nouns with the help of the verbalizer *испытывать*, cf. *радуется – испытывает радость*; *беспокоится – испытывает беспокойство*; *разочаровался – испытал разочарование*. And here again, the combination of *испытывать* with *впечатление* is not accepted by existing standards. The word combination *испытать впечатление* can be met in the Internet. But then *впечатление* should be understood as if it could mean 'feeling', but this meaning of *впечатление* is not acknowledged by the existing dictionaries:

(1.1) Бывают такие *впечатления*, которые не описать словами — их нужно прочувствовать. Они запоминаются навсегда, и *испытать* их хочется каждому (from the Internet);

(1.2) Какие *впечатления* вы *испытали*, встав первый раз на лед в этом году (from the Internet).

Thus, the first puzzle with the word *впечатление* is that there is no derivation model accounting for the semantic relationship of the noun *впечатление* with the verbs *впечатлить, впечатлиться*.

Puzzle 2. **Combinability** of the word *впечатление* is strikingly different from that of typical emotion nouns.

In (Булыгина & Шмелев, 1997) it is said about *удовольствие, радость, огорчение* that they are feelings drawn from the external world, feelings that are d e l i v e r e d to us by the external world. Hence combinability of the words *удовольствие, радость, огорчение* with the verb *доставить* 'deliver'. In fact, something or somebody can "доставить" also *беспокойство, боль, горечь, муки, мученье, наслаждение, неудобство, неудовольствие, отдохновение, страдания, удовлетворение, успокоение* (but not *спокойствие*!).

Still we cannot say **Спектакль доставил впечатление*. Again, the only way to express this idea is to say *Спектакль произвел впечатление*.

It is easy to demonstrate that incompatibility of *впечатление* with *вызвать* should not be treated as the case of idiomaticity, which is accounted for in "Meaning-Text Theory" (Мельчук, 1974) with the help of LEXICAL FUNCTIONS. In fact, incompatibility of *впечатление* with *вызвать* has a clear semantic motivation. Such nouns as *удивление, смущение* belong to the ontological category of states, and states, in fact, can be **caused** by something. Now *впечатление* doesn't combine with *вызвать* because it does not belong to the ontological category of states.

In order to solve puzzle 2 one should ascribe the word *впечатление* (in its different meanings) ontological categories that would predict its combinability and explain its semantic difference from other emotion nouns. That combinability of a word can be accounted for by its ontological category is demonstrated, on the example of emotion nouns, in (Ляшевская & Падучева 2011).

Puzzle 3: **semantic derivation** patterns of the word *впечатление* is different from that of typical emotion words.

A typical emotion word often has several meanings, but it is regular polysemy, usually accounted for by a category change (examples from National Russian Corpus):

(a) *неловкость* – FEELING (Она же *испытывала неловкость*, потому что не могла вспомнить, где она его прежде видела) and EVENT (И я опять не знал, как тут поступить, опять *возникла неловкость*);

(b) *разочарование* – FEELING (Решение доверчивого горсобрания *доставило разочарование* учителям) and STATE (Он положил сверток к ногам летчицы Зины и ушел *в разочаровании*, потому что была у него идея пригласить ее на танцы).

The word *впечатление* comprises a unique combination of meanings, and this fact demands explanation.

Puzzle 4: **argument structure** of the word *впечатление* includes the Genitive explicating the contents of the impression; this Genitive cannot be explained on the basis of the verb *впечатлить*:

(1.3) Их действия, поведение произвели впечатление *отрепетированного спектакля*. [Л.Гурченко. Аплодисменты].

(1.4) Письмо произвело впечатление разорвавшейся бомбы.

In what follows I intend to suggest solutions for all these puzzles.

2 The motivating verb

According to Vasmer's Etymological dictionary of Russian, the word *впечатление* is a loan translation of the French *impression*, which, in its turn, is a loan translation of the German *Eindruck*.

V.V.Vinogradov (Vinogradov, 1994) traces back a longer and more interesting history of *впечатление*. According to V.V.Vinogradov, the noun *впечатление* is derived from the verb *печатлеть* and its derivatives *запечатлеть*, *напечатлеть*, that existed in Old Chirch Slavonic and, afterwards, in Old Russian. V.V.Vinogradov suggests a similar history for *впечатлеть*, which, along with its direct concrete meaning, developed a figurative meaning 'to implement', 'to root'.

The verb *впечатлеть* still existed in Russian in XVIII and the beginning of XIX century; it had the form of imperfective (*впечатлевать*) and the reflexive form *впечатлеться*:

> Через несколько недель получил я ответ – он *впечатлелся* навсегда в моем сердце. (Н.Карамзин. Письма русского путешественника).

There is no doubt that the noun *впечатление* is derived from the verb *впечатлеть*, being its name of RESULT – in the same way as, e.g., *продолжение* 'continuation' (*продолжение следует*) is derived from *продолжить* 'continue'). According to V.V.Vinogradov, the noun *впечатление* was used to convey the meaning of the Latin *impression*; hence convergence with the French *impression* and development of more abstract meanings.

The verb *впечатлеть* gives the clue to puzzle 1 (derivation pattern), puzzle 3 (polysemy, i.e. semantic derivation) and puzzle 4 (argument structure including the genitive of Image).

It is important to bear in mind that in XVIII and early XIX century the word *впечатление* was used not only as the name of result – 'imprint left by the seal', but also as the name of ACTION – 'overlaying the imprint'; in fact, it had regular polysemy normal for nouns with the suffix *-ение*, formed from the verb (Апресян, 1974: 193-203). Example of the word *впечатление* used as nomen actionis:

> Это меня чрезвычайно заняло, и я, для лучшего впечатления этих предметов в моей памяти, вздумал перевести всю статью на русский язык. (Н.И.Греч. Записки о моей жизни.)

Both the meanings of *впечатление* and its relationships with *впечатлеть* are described by V.V.Vinogradov. Still it stands to reason to redo the analysis – with the contemporary linguistic apparatus (elaborated, in particular, in Meaning–Text Theory) and additional material provided by the Russian National Corpus.

3 Diatheses of the motivating verb

Examples with the verbs *впечатлеть*, *впечатлеться* from the XVIII and early XIX century:

(3.1) Боже отмщений! Тако ли и я казнюся, как был казним Каин? *Впечатлел* ли ты на челе моем знаки моего злодеяния? [Д. И. Фонвизин. Иосиф (1769)]

(3.2) Пан Меховецкий, друг первого обманщика, сделался руководителем и наставником второго; *впечатлел ему в память* все обстоятельства и случаи Лжедимитриевой истории, [Н. М. Карамзин. История государства Российского: Том 12 (1824-1826)]

(3.3) Лишь бы только мрачная злоба людей *не впечатлела* <…> *в мягкое его сердце*

недоверчивости, ненависти к людям (А.И.Тургенев; цит. по ССРЛЯ)

(3.4) Творец <…> *впечатлел в нем* [человеке] *образ* и подобие свое [архиепископ Платон (Левшин). Слово на день Рождества Пресвятыя Богородицы (1780)]

(3.5) Сия картина так сильно *впечатлелась* в его юной душе, что он через двадцать лет после того, не мог без особливого радостного движения видеть большой реки, плывущих судов, летающих рыболовов. [Н. М. Карамзин. Рыцарь нашего времени (1803)]

(3.6) Великодушная государыня ужаснулась и <…> произнесла слова, которые хотя не могли перейти к нам во всей точности, но глубокий смысл их *впечатлелся* в сердцах многих. [Н. В. Гоголь. Портрет (1835)]

3.1 Agentive diathesis of *впечатлеть*

A *впечатлел на* Y-*е образ* Z X-*а* = 'A created on Y an imprint Z of X'.

The verb *впечатлеть* denotes a situation with four participants: A is the Agent, Z is the Imprint of the Prototype X on the Recipient Y. For example:

(3.1) = *Впечатлел* ли ты (A) на челе (Y) моем знаки (Z) моего злодеяния (X)?

Participant Y can be marked both by Prepositional case, as in (3.1), and by Accusative, see (3.2) *впечатлел ему в память все обстоятельства*.

Thus, an impression is not d e l i v e r e d to Y from the outside (as *радость* 'joy'), it is not caused (as *тревога* 'trouble'), it is i m p r i n t e d into / in Y in the form of Z. This imprinting can be produced by an Agent A, as in example (3.1). But the participation of the Agent is optional: X can imprint into Y by itself, producing its image (= impression) Z.

3.2 Non-agentive diathesis of *впечатлеть*

X *впечатлел в* Y <свой> *образ* Z = 'X created on Y its imprint Z'.

The domain of values of the argument Y becomes more narrow: usually, Y is *душа, сердце, память* of a person; not even *чело* 'forehead', as in (3.1). The place of the Agent is occupied by X, but participant Z preserves its role of the Imprint (Image) of X:

(3.3) = Лишь бы только мрачная злоба людей (X) *не впечатлела* <…> *в мягкое его сердце* (Y) *недоверчивости* (Z) [here *недоверчивость* Z is the imprint of *злоба*, perhaps, figuratively: literally, *недоверчивость* is, rather, the result of influence of *злоба*)].

4 Participant Imprint (Z) and valence for Z

In its initial meaning the word *впечатление* is the name of RESULT derived from the verb *впечатлеть*; in terms of Meaning–Text theory, it is the name of the 2d argument of the verb *впечатлеть* (S_2):

впечатление = 'то, что A (или сам X) впечатлел в Y-е /в Y как образ X-а'.

In other words, *впечатление* is an imprint of X in /into Y. The literal meaning 'stamp' is preserved as late as in texts of N.M.Karamzin:

(4.1) Красота Лизы при первой встрече сделала *впечатление* в его сердце. [Н. М. Карамзин. Бедная Лиза (1792)]

In accordance with the rules distinctly formulated in MTT, the valence for Z, obligatory for the verb *впечатлеть*, disappears in *впечатление* as S_2: it is supposed to be filled by the

noun itself.

But in some contexts it is possible to treat *впечатление* as a name of action, nomen actionis. In fact, *произвести впечатление* can be understood in the same way as *произвести проверку*, i.e. *проверить* 'to check'. In the first case, when *впечатление* is the name of result, the valence for Z disappears, in the second case it preserves:

A *впечатлел на /в* Y-*е* / *в* Y *образ* Z X-*а* =

A *произвел на/в* Y-*е или в* Y *впечатление* X-*а в виде* Z-*а*.

Thus, relationships between

впечатлеть = *произвести впечатление* = *создать впечатление*

are the same as between

вклеить = *произвести* [i.e. 'carry out'] *вклейку* = *создать* [i.e. 'make'] *вклейку*.

The verb *впечатлеть* belongs to the class of IMAGE CREATION verbs (Levin 1993: 169, Апресян 1996, Падучева 2003). These verbs imply participants Image and Prototype. For example, one can paint *a general* and a portrait of a general; one can *впечатлеть*, on the one hand, *пустыню мрачную*, as in (4.2a), and *образ*, as in (4.2b)

(4.2) a. Ты живо *впечатлел* в моем воображеньи Пустыню мрачную, поэта заточенье, Туманный свод небес, обычные снега И краткой теплотой согретые луга. [А.С.Пушкин. К Овидию (1821)]

b. Пребудет *образ* ввек во мне, Она который *впечатлела*! [Г. Р. Державин. (1783-1784)]

Thus, we have got the solution for puzzle 4, namely, of the Genitive in the example (1.4) with a bombshell – this Genitive fills the valence for Z:

(1.4) Письмо произвело впечатление разорвавшейся бомбы = 'the letter imprinted <in the consciousness> the image of a bombshell'.

Now let's go down to the meanings of the word *впечатление* in modern Russian.

5 Meanings of the word *впечатление*

Dictionaries differentiate three meanings of *впечатление*, which have no overt connection between one another. These connections become transparent if we begin with the verb *впечатлеть*. Below each meaning is provided with an ontological category.

5.1 Meaning 1 (*впечатление* is an IMAGE)

***впечатление от* X-*а у* Y-*а* = 'image (Z) that X imprinted in the consciousness of Y'**.

This meaning is represented by examples (5.1) – (5.6); participant Y can be implied to be a collective consciousness, not the consciousness of an individual, so argument Y is omitted. The valence for Z, image, is cancelled, because *впечатление* is itself the image.

(5.1) общее *впечатление от команды* становится целостным и радующим глаз. [«Известия», 2003.02.09]

(5.2) <…> который, впрочем, *впечатления* от "Спартака" не *попортил*, а только напомнил, кто в избе хозяин. [«Известия», 2002.10.23]

(5.3) Я помню себя рано, но первые мои *впечатления* разрозненны (В.Г.Короленко, цит. По МАС).

(5.4) *Впечатления* от похорон могут вызвать серьёзный регресс в развитии ребёнка. [//«Домовой», 2002.08.04] [*впечатления* = 'imprints in consciousness of what a

child had seen at the funeral', 'recollections']

(5.5) ей нравилась ранняя утренняя пустота Москвы, даже ноябрьская мокрая мгла не портила *впечатления*. [Анна Берсенева. Полет над разлукой (2003-2005)] [*не портила* = *впечатления* = 'didn't spoil the picture in the consciousness']

(5.6) Незабываемое *впечатление* осталось от голоса Анны Литвиненко. [«Российская музыкальная газета», 2003.04.09] [it is the image in the consciousness that is unforgettable']

Thus, the outside world doesn't "deliver" the impression to our consciousness – it literally c r e a t e s the impression; in fact, *впечатлеть* is a verb of creation.

Meaning 1 is the most material of the meanings of our word: *впечатление* 1 is a kind of stamp. Semantically, the word combination *впечатление от спектакля у ребенка* 'child's (Y) impression of the performance (X)' exploits the same pattern as, e.g., *изображение храма на гобелене* 'image of the temple (X) on the tapestry (Y)':

Но с годами флер очарования его героизмом и писательским талантом рассеялся, и мои *впечатления отфильтровались* во вполне четкую картину. [Нина Воронель. Без прикрас. Воспоминания (1975-2003)]

5.2 Meaning 2 (*впечатление* is an IMPACT <ВОЗДЕЙСТВИЕ>)

This meaning occurs, in the first place in the context of the verb *произвести* 'produce':

X *произвел на* Y *впечатление* Z-*а* = 'X incorporated in the consciousness of Y its image Z'.

NB the valence for the participant Z. Examples:

Молодая, приветливая блондинка, веселая хохотушка, она (X) *произвела* на меня *впечатление* совершенно несерьезной женщины (Z). [В.Запашный. Риск. Борьба. Любовь (1998-2004)]

Он (X) *производил впечатление* очень интеллигентного, обременённого жизненным опытом и глубокомысленного человека (Z). [LiveJournal (2004)] = 'X отражался в сознании людей как Z'

The contents of the impact can be conveyed by an adjective:

На Андре Жида архитектура Москвы произвела *удручающее* впечатление [«Неприкосновенный запас», 2003.07.14]

Вид нежилого неразрушенного города произвел *гнетущее* впечатление. [«Искусство кино», 2003.06.30]

But it can be the case that the impact is characterized only from the point of view of its existence and strength, while the contents of the impact remains undisclosed, i.e. participant Z is off stage:

Лизавета Ивановна *произвела впечатление* на майора;

Деловитость *произвела впечатление* на Зимина, и он взял в свою команду нового человека. [«Знание – сила», 2003]

The lexeme *впечатление* 2 has another diathesis:

Y *находится под впечатлением от* X-*а* = 'Y is in mental or emotional state engendered by the pressure of X'.

<Из этого Милий Алексеевич заключил, что> *майор* всё ещё *под впечатлением от* Лизаветы Ивановны. [Юрий Давыдов. Синие тюльпаны (1988-1989)]

На следующий день *я* гулял по берегу моря, всё ещё находясь *под впечатлением свидания*, вспоминая его волнующие подробности и, главное, чувствуя себя на

голову выше, чем до него. [Фазиль Искандер. Письмо (1969)]

Double diathesis of *впечатление* 2 generates the relationship of conversion, e.g., between (a) and (b):

(a) Ваша деловитость *произвела впечатление* на директора

(b) Директор *находится под впечатлением* от вашей деловитости,

Note that in this context participant Z is excluded, which fact is suspicious and demands explanation.

5.3 Meaning 3 (*впечатление* is an OPINION):

<*от* X-*а*> у Y-а *впечатление*, что Z = '<observing X> Y came to the opinion that Z'.

In this meaning the participant Z, the opinion, is obligatory; note the possibility of governing the subordinate *that*-clause, typical for words of opinion, both verbs and nouns. X is a situation observed by Y which was the source of the impression Z, in this case conveyed by a proposition. Participant X may remain unexpressed.

У Тани *создалось впечатление, что* они играют в какую-то взрослую игру — делят что-то понарошку… Но делили взаправду… [Л.Улицкая. Путешествие в седьмую сторону света (2000)]

А у меня *осталось впечатление, что* упущена реальная возможность предотвратить серьезную ошибку. [Г.Арбатов. Человек Системы (2002)]

У меня *впечатление, что* я телеграфирую в пустоту. [Л.Смирнова. Моя любовь (1997)]

6 Dynamic semantics of the word *впечатление*

Thus, the word *впечатление* has three meanings, each of which has its own hyponym, i.e. ontological category: IMAGE, ACTION, OPINION. It follows, then, that *впечатление* doesn't belong to any of ontological categories typical for emotion words according to (Ляшевская & Падучева 2011) – such as STATE, RELATION or FEELING. There is no other verbal noun with such combination of meanings. But, at least, the verb *впечатлеть* makes it possible to derive these meanings from a single source applying general derivation rules.

Meanings 1 and 2 descend directly from the verb *впечатлеть*, meaning 1 being the name of result, while meaning 2 is the name of action. Transition from meaning 1 to meaning 3 can be presented as a metaphoric shift, i.e. a change of concept: it is a transition from an image of a situation to a proposition describing it.

As a rule, the three meanings of *впечатление* are sharply differentiated by the context. Ambiguity may arise in the context of indirect question; for example, in (6.1) *впечатление* is understood as 'impact', while in (6.2) it can be both 'impact' and 'image':

(6.1) Легко вообразить, какое <большое> *впечатление* Алексей должен был произвести в кругу наших барышень. [А. С. Пушкин. Барышня-крестьянка (1830)]

(6.2) Вы, надеюсь, понимаете, какое *впечатление* он на меня произвёл. [Анатолий Рыбаков. Тяжелый песок (1975-1977)]

Example (6.3) is of fundamental importance, for *впечатление* is used here in its two meanings simultaneously – *вынес впечатление* 'got an impression' semantically agrees with 'image', while *огромное впечатление* 'great impression' concords with 'impact':

(6.3) Я помню, какое огромное *впечатление вынес* от этого произведения. [«Вестник

США», 2003.10.29]

In fact, According to Roman Jakobson, simultaneous appeal to the two different meanings of a word provides the effect of a pun; but for metonymically related meanings such a conflation is possible (Падучева, 2004: 416).

All the meanings of the Russian *впечатление* (in contradistinction to its translational equivalents in English, German or French) belong to the sphere of the ideal, though the association of the inner form of *впечатление* with *печать* is clear for a native speaker of Russian:

На какой «внутренний субстрат» «печатается» то, что оказывает на нас впечатление? – задает вопрос психолог, изучающий связь между впечатлениями и эмоциями [RNC].

Conclusion

To recapitulate, the obsolete verb *впечатлеть* made it possible:

- to reveal word formation patterns that connect the verbal noun *впечатление* with the verb;
- to discern relationships between the meanings of the noun *впечатление*;
- to describe combinability of the word *впечатление* as motivated by its ontological categories;
- to explain specific diathesis of the noun *впечатление* comprising the Genitive of Image.

Examples (1.1), (1.2) show, however, that the word *впечатление* experiences pressure from its neighbors in the thematic class (of emotions) and by and by acquires combinability inherent for prototypical nouns of emotion, namely, for nouns of state.

The word *впечатление* is an example of the following important lexicographic phenomenon. Two of its meanings are generated by a productive derivational pattern but from a word that doesn't exist in modern language. This phenomenon has now become an object of attention in lexical semantics: productive derivatives of extinguished meanings are discussed, e.g., in (Бабаева, 1998), (Урысон, 2005). It is a fruitful field of exploration.

Bibliography

Апресян, Ю.Д. 1974. *Лексическая семантика: Синонимические средства языка*, Москва: Наука.

Апресян, Ю. Д. 1996. О толковом словаре управления и сочетаемости русского глагола. In *Словарь, грамматика, текст: Сборник в честь Н. Ю. Шведовой*, 13–43, Москва: Наука.

Бабаева, Е.Э. 1998. Кто живет в вертепе, или Опыт построения семантической истории слова, *Вопросы языкознания* №3, 94–106.

БАС – Словарь русского литературного языка. В 17 т. Москва-Ленинград: Изд-во АН СССР, 1950–1965. («Большой» академический словарь).

Булыгина Т.В. & Шмелев А.Д. 2000, Перемещение в пространстве как метафора эмоций, In *Логический анализ языка. Языки пространств.* Москва: Индрик.

Виноградов В.В. 1994. История слов. Москва: ТОЛК.

Ляшевская О.Н.& Е.В.Падучева, 2011. Онтологические категории имен эмоций, *Научно-техническая информация* №5, http://lexicograph.ruslang.ru/05News.htm .

МАС – Словарь русского языка. В 4 т. Москва: Русский язык, 1981. («Малый» академический словарь)

Мельчук И.А., 1974. *Опыт теории лингвистических моделей «Смысл – Текст». Ч. 1. Семантика, синтаксис*. Москва: Наука; издание 2-е Москва: Языки русской культуры, 1999.

Падучева Е.В. 2003. Глаголы создания образа: лексическое значение и семантическая деривация. *Вопросы языкознания* № 6, 30-46.

Падучева Е.В., 2004. *Динамические модели в семантике лексики*. Москва: Языки славянской культуры.

Урысон Е.В., 2005. Логическая структура полисемии и ее реализации (слово СЛЯКОТЬ системе языка), *Русский язык в научном освещении*. 2 (10): 87-120.

Levin B., 1993. *English Verb Classes and Alternations: A preliminary investigation*. Chicago: Chicago University Press.

Measurement-Contents Constructions in Russian

Olga Jurievna Podlesskaja

ITTP RAS
olga@iitp.ru

Abstract

This paper is dealing with a special subclass of Russian prepositional idiomatic constructions for naïve measurement with preposition ***v*** and a noun in accusative case – *dom v tri etaža* 'a house three storeys high', *sinjak v pol-lica* 'a black eye as big as half of the face', *kover vo vsju stenu* 'a carpet as big as a whole wall'. They are semantically different depending on the participants of the construction (numerals, ***ves'*** 'whole' and ***pol-*** 'half') – such as what is being measured, the object or the part of the object and what is being used as a measure. The construction with numeral may also be called contents construction. It is shown that this type is easily paraphrased in construction with an attribute. Also it turns out that constructions with modifiers ***ves'*** 'whole' and ***pol-*** 'half' despite their inner semantic differences have the same meaning of big size or high degree. Some semantic and syntactic peculiarities of these constructions are described.

Keywords

Syntactic phrasemes, measurement constructions, naïve measurement, semantics

1 Introduction

There are interesting idiomatic prepositional constructions in Russian with prepositions *v* 'in' and *s* 'with' that require accusative case after them: *rybka razmerom v ladon'* 'a fish as big as a palm' / *mal'čik rostom s pal'čik* 'a boy as big as a finger'. They represent a naïve measurement of objects, independent from common stereotypical means of exact measurement. In this system the objects correspond directly to the speaker's experience – by comparing the sizes with the sizes of human body parts and other objects, relevant to the speaker and used as standards for size and form [Shemanaeva 2008]. All components of this construction, such as object, measure, and parameter of measurement (length, height, width, etc.) are bound with each other and depend on the choice of the other parts.

The fact that constituents of the construction semantically depend on each other leads to the conclusion that we are dealing with special units on the boundary between lexicon and grammar. They are called constructions by Charles Fillmore, see (Fillmore et al., 1988),

(Goldberg 1995), (Jackendoff 1997), (Kopotev 2008), (Plungian & Rakhilina 1996), (Podlesskaya & Rakhilina 2000), (Podlesskaya 2007), units of microsyntax, or syntactic phrasemes, in Moscow Semantic School (Iomdin 2003, 2006, 2008), and a type of phrasemes in (Melčuk & Iordanskaja, 2007).

The subject of this paper is a special sub-class of prepositional constructions with preposition v 'in' and accusative: *dom v tri okna* lit. 'house in three windows' = 'a house three windows wide / a house that has three windows', *komnata v pol-etaža* lit. 'a room in half-storey' = 'a room as big as half a storey / a room occupying half of the floor', *pjatno vo vsju stenu* lit. 'a stain in whole wall' = 'a stain as big as a whole wall / a stain the size of a wall', and synonimic attributive constructions: *trexetažny dom* 'three-storey house', *pjatiaktny balet* 'ballet in five acts' and so on. The goal of the description is to define the most probable and relevant participants of this situation of measurement, the parameters connected to the defined participants, some restrictions and the general meaning of the whole construction.

2 Comparing with the Outer Measure vs. Measuring Using Inner Parts

All prepositional measuring idiomatic constructions with preposition ***v*** 'in' and accusative, which were mentioned above, have the same structural appearance. However, despite this structural similarity, they denote three semantic situations, three types of relations between the measure and the measured object:

1. the measure comes from relevant measuring domain, but is **not** a part of the measured object – *mal'čik rostom v pal`čik* 'a boy as big as a finger' (not 'as big as his own finger'), *rybka v ladon'* 'a fish as big as a hand', *sloj pyli tolščinoy v tri pal'ca* 'a dust layer as thick as three fingers';

2. the measure is an important (focused) part of the object – as in *dom v tri etaža* 'a house three storeys tall';

3. the object is a part of the measure – as in *pjatno vo vsju stenu* 'a stain as big as a whole wall', *komnata v pol-etaža* 'a room as big as half a storey'.

Case 1) may be illustrated with the following scheme of measurement constructions, see Table 1:

X (measure object)	V (copula)	Y (parameter in instrumental case)	(Preposition ***v***	Z (measure in accusative case)
Mal'čik	*Byl*	*Rostom*	*V*	*pal'čik*
Rybka			*V*	*ladon'*
Sloj pyli		*Tolščinoj*	*V*	*3 pal'ca*

Table 1: Scheme of measurement constructions

Some participants are obligatory, some may be explicitly omitted; the construction may convey both big and small size, depending on the arguments [Shemanaeva 2008]

In this paper we focus on cases 2) and 3), which represent two possible measurement strategies within one situation frame where object and measure are connected by the part-whole relationship.

3 Measuring Using Parts of the Object: Contents or Sizes ?

When we measure the object by its parts, it is difficult to say that we actually measure it because first we state the contents and the number of constituents and then we indirectly convey also the information of the size of the object. To state the number of constituents we need a numeral in the construction: *lestnica v odin prolet* 'a staircase with one flight', *dom v tri etaža* 'a house three storeys high', *stado v tysjaču golov* 'a flock/ a herd of one thousand heads' and so on. It can be viewed as a measurement because implicitly *dom v tri etaža* 'a house three storeys high' is bigger than *dom v odin etaž* 'a house one storey high', so we judge whether an object is relatively big or small by the number of its relevant parts.

Interestingly, if an object has more than one parameter of measuring (more than one relevant dimension of measuring), there is usually a dependence between the part and the parameter: the house in *dom v tri okna* 'a house three windows wide / a house that has three windows' is measured only in horizontal dimension (despite the fact that the windows can also be theoretically counted vertically like floors), while the house in *dom v tri etaža* 'a house three storeys tall' is inevitably measured in vertical dimension.

4 "Close Contact" Measurement: Part vs. Whole

The other possible measurement strategy within the situation frame where object and measure are connected by part-whole relationship is when we denote the size of the object not by its smaller constituents and their number, but by the comparison to the whole, which turns out to be a measure.

This is the case of prepositional idiomatic constructions with *v* and modifiers: *ves'* 'whole' and *pol-* 'half': *v pol-lica* 'as big as half the face', *vo vsju stenu* 'as big as the whole wall'. The whole is a measure and its part is being measured: *šram v pol-lica* 'a scar as big as half the face', *zerkalo vo vsju stenu* 'a mirror as big as the whole wall'. The part is situated on the whole, the location is one of close contact, either physically like in *šram v pol-lica* 'scar the size of half the face', or visually, like in *očki v pol-lica* 'glasses the size of half the face'.

In spite of the fact that only a half of the whole measure is mentioned, the construction ***v pol-Z*** itself has a meaning of high degree or big size, as if the measure for the size was whole and thus big, much bigger that the part itself: *rumjanec v pol-lica* ≈ 'bright flush', *šram v pol-lica* ≈ 'big scar', cf. also *glaza / očki / rot / fingal (sinjak) / pjatno v pol-lica* 'eyes / glasses / mouth / black eye / stain as big as half the face', *fortočka / cvetok v pol-okna* 'a ventlight / flower as big as half the window'.

The size of the part of the measure, normally relatively small, through the metaphorical comparison with the size of the whole (*vo vsju stenu* 'as a whole wall') and even to half of the size of the whole (*v pol-okna* 'as half the window') is thus regarded as very big.

We observe that constructions with both modifiers – *ves' and pol-* denote the same meaning of big size or high degree. This situation has a semantic parallel – the case of Russian *polu-*, which can mean either half[1] of the thing or state (*polupustoj* 'almost empty' – *pustoj* 'empty') or the whole thing (*polumesjac* 'moon' – *mesjac* 'moon'), see (Iomdin 2003).

Below, we will look more closely at the structure of the two constructions ***X vo ves' Z*** and ***X v pol-Z***.

X, as we stated above, should be a relevant part of Z, being in either physical or visual contact with Z. As for Z, these objects tend to be 1) flat spacious surfaces (*stena* 'wall', *okno* 'window', *ekran* 'screen'), also the surfaces on human body, that is to say, body parts that are conceptualized as surfaces (*lico* 'face', *ščeka* 'cheek', *spina* 'back', *grud'* 'chest'[2]); or 2) spaces (*komnata* 'room', *zal* 'hall', *ulica* 'street', *ploščad'* 'square', *gorod* 'city', *nebo* 'sky', *zemlja* 'earth'). The measures apart from their topological characteristics should be relatively big, cf. *??? carapina vo ves' palec* 'a scar as big as a whole finger'. The measures may also be "unmeasurable", but they are used in the naïve measurement nevertheless: *raduga v pol-neba* 'a rainbow as big as half the sky', *grjada voln vo vse more* 'ridge of waves as wide as the sea', and

> *Vnizu ležala širokaja, v polzemli, zelenaja polosa…* [I. Grekova]
> 'Below lay a wide, as wide as half the earth, green patch'

The parameter **Y** (general size, length, width) may be explicitly expressed with an adjective denoting size (*ogromnyi* 'huge', *dlinnyi* 'long', *bol'šoj* 'big') like in schemes ***Adj X v pol-Z*** or ***Adj X vo ves Z:*** *gromadnaja kletka v polkomnaty* 'a huge cage as big as half the room', *bol'šaja, v polsteny, fotografija* 'a big photo as big as half the wall'. The construction itself has a meaning of big size (*kletka v polkomnaty,* a cage as big as half of the room is evidently big) so this repetition of expressive means is emphatic. If the parameter denotes small size, it does not go well along with the construction of generally big size and high degree. An expression *nebol'šie štorki v pol-okna* 'not very big curtains as big as half the window' shows some conflict between the semantics of the whole construction and the semantics of its constituents.

The parameter Y may also be in instrumental case, as shown in the table above for other ***v-*** constructions. In this case it is important to distinguish between the approximative measurement, as in our constructions, and the exact measurement, which is of no big interest to us because the scope of compatibility and acceptability of objects and measures is wider, such as in *širinoj v dva metra* 'as wide as two meters'*, dlinoj v pol-šaga* 'as long as half the

[1] Or even the absence of the state mentioned, cf. *poluodetyi* – 'almost nude, not fully clothed'.

[2] *Spina* 'back' and *grud'* 'chest' are often used metonymically as corresponding parts of the clothes, so there are many examples like *pjatno vo vsju grud'* 'stain as big as a whole chest', *vyšivka vo vsju spinu* 'embroidery as big as a whole back'.

step', and so on. The idiomatic construction with omitted parameter and exact measure is not so liberal: *[???]kartina v dva metra* 'a picture as big as two metres', *[???]šram v tri santimetra* 'a scar as big as three centimeters'.

The parameter may also have a place of Z in the ***ves'***-construction: ***X vo ves Y Z-a***, like in *vo vsju širinu ulicy* 'as wide as the width of the street' or *vo vsju dlinu komnaty* 'as long as the length of the room'.

The construction ***X vo ves' Z*** (with certain Z such as parameters *dlina* 'length', *šir' / širina* 'width' and some others) can be either an adverbial modifier[3] or an attributive group for a noun, cf.

> *stojali vo vsju dlinu tanceval'noj zaly* 'were standing along the length of dance hall' vs.
> *stol vo vsju dlinu tanceval'noj zaly* 'a table as long as the length of dance hall'

We will show some illustrations of verb and noun modifiers in context (though the examples are taken from fiction and they are not so widely spread in colloquial language):

Verb modifiers:

A molodoj odnonogij elektrik Perepelicyn medlenno ***pokrasnel vo vsju ščeku*** **'flushed to the width of a whole cheek'** *i ne skazal ni slova...* [V. Grossman][4]

Fedjunja opjat' uvernulsja, Kirsha i ***pljuxnulsja vo vsju spinu*** **'fell down to the length of the whole back'**, *a Fedjunja tut kak tut…* [P.P. Bažov]

On protjanul ruku za sigaretami, ščelknul zažigalkoy, gluboko, ***vo vsju grud' zatjanulsja*** **'inhaled as deep as a whole chest'**. [G. Baklanov]

Noun attribute:

Udivitel'no xoroš, krasavec, možno skazat'. Strojnyj, vysokij, ***rumjanec vo vsju ščeku*** **'flush as big as a whole cheek'**... [A. S. Pushkin]

V konce koncov mne bylo veleno pereodet'sja v temnyi paradnyi kostjum xozjaina nomera s **nagradami vo vsju grud' 'awards as big as a whole chest'**. [B. Griščenko]

When the parameter Y is omitted and the object may be measured in various directions, we may have difficulties in understanding. *Vo vsju ulicu* 'as big as a whole street' is ambiguous because *ulica* 'street' has both width and length, whereas *transparant vo vsju ulicu* 'banner as big as the whole street' or *očered vo vsju ulicu* 'a queue as long as the length of the street' are not ambiguous. The expression *kover vo vsju komnatu* 'a carpet / a rug as big as a whole room' is not quite understandable because the carpet may be either on the floor or on the

[3] A lot of Russian idiomatic prepositional constructions with ***v*** and noun in accusative case do not denote size at all but they denote high degree and appear only as adverbial modifiers, cf. *vo vse gorlo* 'by full throat' */ vo vsju moč* 'with all one's might' */ vo ves' dux* 'with all one's might' */ vo vsju pryt'* 'at full speed' and so on.

[4] Most examples were taken from the National Corpus of Russian Language (www.ruscorpora.ru).

wall, so it may be measured in horizontal or vertical dimension. Only from the context we may guess that the carpet is actually lying on the floor:

Verojatno, tut byl ***kover vo vsju komnatu****, potomu čto, edva sdelav dva ili tri šaga, on kuda-to propal, a potom takie že šagi razdalis' v protivopoložnom konce etix potemok.* [B.L. Pasternak]
'There was probably a carpet as big as a whole room here because he disappeared after two or three steps and then the sound of his footsteps was heard in the opposite corner of this dark room'

The modifier (***ves'*** or ***pol-***) is an obligatory participant of the construction (**šram v lico* 'a scar as big as a face', **kover v stenu* 'a carpet as big as a wall') and the only possible attribute in it, cf. *[???]vo vsju nekrašenuju stenu* 'as big as an unpainted wall', *[???]vo vse nemytoje okno* 'as big as an unwashed window', *[???]vo vsju puxluju ščeku* 'as big as a plump cheek', *[???]vo vse detskoje lico* 'as big as a child's face'. Exceptions are possessive pronouns: *vo vsju našu stenu* 'as big as our wall', *vo vse moe okno* 'as big as my window'[5].

As for the choice of the participants in the measurement construction, we may state that some Zs tend to appear mostly in ***ves'*** construction (*vo vsju grud*'), some others mostly in ***pol-*** construction (*v pol-lica*), while still others combine with both (*v pol-steny – vo vsju stenu*).

5 Attributive Measurement Construction

We can create a certain hierarchy of objects that are used as measures: (1) objects that belong to the whole, its parts (a house has floors, a staircase has flights); (2) abstract measurement units (meters, centimeters); (3) standard objects that do not belong to measured possessor (*rybka v ladon'* 'a fish as big as a hand', *sloj pyli tolščinoy v tri pal'ca* 'a dust layer as thick as three fingers').

Abstract measurement units are most easily paraphrased into attributive constructions (*volna v tri metra* 'a wave as big as three meters' – *trexmetrovaja volna* 'three meter wave'), while it is slightly more difficult for parts of the whole (*trexproletnaja lestnica* 'three-flighted staircase') and it is almost unacceptable for body parts as measurement units. The corresponding adjectives (*trexpalyi* 'three-fingered', *dvuxgolovyi* 'two-headed') denote that the object has a certain number of these body parts and not the size of this object.

Paraphrasing is not acceptable either for the constructions with ***ves'*** and ***pol-***, where a part of the object is measured by being compared to the whole. *Stennoj škaf* lit. 'wall wardrobe' does not mean that it is as big as a wall neither does *komnatnoje rastenije* 'room plant' tell us anything about its size. Components ***ves'*** and ***pol-*** do not compose compound adjectives in paraphrasing constructions at all.

[5] In the measurement construction with preposition ***s*** 'with' and accusative case this situation is common: *kulak razmerom s detskuju golovu* 'a fist as big as a child's head', *jabloko s horošij arbuz* 'an apple as big as a good watermelon'.

6 Conclusion

We have shown in this paper that there exist different types of naïve measurement expressed in prepositional constructions with the similar structure. The measurement depends on what object is being measured – a bigger object consisting of small parts or parts of the bigger objects. The two types – contents and size construction have different abilities of being paraphrased into attributive constructions.

The measurement idiomatic construction with preposition ***v*** and accusative denoting big size or high degree has its own semantics. It correlates with the fact that the component *pol-*, or *polovina,* 'half' does not necessarily mean that the size of the measured object is small. Thus, constructions with half the measure and the whole measure appear to be nearly quasi-synonymic, while a part and the whole are not quasi-synonyms: *kletka vo vsju komnatu* 'a cage as big as a whole room'/ *kletka v pol-komnaty* 'a cage as big as half the room'.

Bibliography

Fillmore, Ch., P. Kay, & M. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone. Language*, 63(3): 501-38.

Goldberg, A. 1995. *Constructions: A constructionist grammar approach to argument structure.* Chicago: Chicago University Press.

Jackendoff, R. 1997. Twisting the night away. *Language*, 73(3): 535-559.

Kopotev, M. 2008. *Principy sintaksičeskoy idiomatizacii*. Helsinki.

Iomdin, B.L. 2003. Semantika russkoy pristavki POLU-. In *Rusistika na poroge XXI veka: problemy i perspektivy. Materialy meždunarodnoj naučnoj konferencii.* Moscow, IRL RAS, pp. 109–113.

Iomdin, L.L. 2003. Bolšie problemy malogo sintaksisa. In *Trudy meždunarodnoy konferencii po kompjuternoj lingvistike i intellektual'nym texnologijam. Dialog'2003*. Moscow: Nauka, pp. 216-222.

Iomdin L.L. 2006. Mnogoznačnye sintaksičeskie frazemy: meždu leksikoy i sintaksisom. In Kobozeva I.M., Narinyani A.S., Selegey V.P. (eds.) *Kompjuternaja lingvistika i intellektual'nye texnologii: Trudy meždunarodnoy konferencii Dialog'2006.* Moscow: Nauka, pp. 202-206.

Iomdin, L.L. 2008. V glubinax mikrosintaksisa: odin leksičeskij klass sintaksičeskix frazem. In *Kompjuternaja lingvistika i intellektual'nye texnologii: Trudy meždunarodnoj konferencii Dialog'2008*. Issue 7(14). Moscow: RGGU, pp. 178-184.

Melčuk, I. & L. Iordanskaja. 2007. *Smysl i sočetaemost' v slovare*. Moscow: Languages of Slavic cultures.

Plungian V.A., Rakhilina E.V. 1996. "Tušat-tušat – ne potušat": grammatika odnoj glagol'noj konstrukcii In Zmarzer V., Petrukhina E.P. (eds.) *Issledovanija po glagolu v slavjanskix*

yazykax: glagol'naja leksika s točki zrenija semantiki, slovoobrazovanija, grammatiki. Moscow : Filologiya.

Podlesskaya V.I. & Rakhilina E.V. 2000. Licom k licu. In Arutyunova, N.D., Levontina, I.B. (eds). *Jazyki prostranstv. Logičeskij analiz jazyka.* Moscow: Languages of Russian culture.

Podlesskaya V.I. 2007. Mnogoznačnost konstrukcii «č*to* + *za* + NP» v svete dannyx NKRYa: čto že eto za konstrukcija?! In Iomdin L.L., Laufer N.I., Narinyani A.S., & Selegey V.P. (eds.) *Kompjuternaja lingvistika i intellektual'nye texnologii: Trudy meždunarodnoj konferencii Dialog'2007*. Moscow: RGGU, pp. 460-469.

Shemanaeva O.Yu. 2008. Konstrukcii razmera v tipologičeskoj perspektive. PhD thesis. Moscow.

Russian Nouns with "Voice Sound" Meaning: Relations of Semantics and the Paradigm of Number

A. Ptentsova

Moscow State University
Moscow, Novoorlovskaya str., 8-2-120
anna.ptentsova@gmail.com

Abstract

Nouns denoting voice sounds can mean single or multiple, short or prolonged sounding; they can imply one or many sound producers. These specific features result in possibility / impossibility to create the plural form or the meaning of this form by such nouns.

The Russian nouns GALDEŽ 'hubbub, din' (in its 2 meanings), GAM 'clamour', GVALT 'uproar', GOMON 'hubbub', OR 'shouting', KRIK 'cry', 'shout' (in its 3 meanings) and VOPL' 'scream', 'yell' are concidered in the paper in terms of the above mentioned criteria.

Keywords

semantics, polysemy, paradigm of number.

The paper states some observations made by the author while compiling dictionary entries of the lexemes under consideration for the Russian Active Vocabulary Dictionary.

Nouns denoting voice sounds as well as other names of sounds can mean single or multiple, short or prolonged sounding; they can imply one or many sound producers. These specific features result in possibility / impossibility to create the plural form or the meaning of this form by such nouns.

Let us consider in this respect Russian nouns **GALDEŽ 'hubbub, din'** (in its 2 meanings), **GAM 'clamour'**, **GVALT 'uproar'**, **GOMON 'hubbub'**, **OR 'shouting'**, **KRIK 'cry', 'shout'** (in its 3 meanings) and **VOPL' 'scream', 'yell'**.

On the whole, the names of sounds can be classified into 2 groups: the first group comprises names denoting homogeneous or a multiple repeated sound. Such sound is understood as indiscrete and the related nouns belong to singularia tantum type. Compare: *plesk 'splash', šelest 'rustle', ščebet 'twitter', lepet 'babble', rokot 'roar', šepot 'whisper', smex 'laughter', xoxot 'loud laughter'* and others.

The second group presents countable singular nouns with the full paradigm of number; such nouns denote discrete sounds. Compare: *gudok-gudki 'whistle' – 'whistles', zvonok – zvonki 'bell' – 'bells', smešok – smeški 'giggle' – 'giggles', xlopok – xlopki 'clap'– 'claps'* and others [1].

From the above mentioned names of voice sounds the nouns **GALDEŽ**, **GAM**, **GVALT**, **GOMON**, **OR** belong to the first group. When it comes to the nouns **KRIK** and **VOPL'**, they can possess the properties of the both groups as it is to be shown later.

Let us consider the nouns of the first type.

GALDEŽ, **GAM**, **GVALT**, **GOMON** usually denote prolonged sound. Compare: *postoyanny<beskonečny > galdež <gam, gvalt, gomon> 'constant <endless> hubbub <clamour, uproar, din>'* VS quite impossible combinations as **nedolgiy, *korotkiy galdež <gam, gvalt, gomon>; *minutny gam *'short<brief> hubbub <clamour, uproar, din>; *'a minute clamour'* and quite doubtful combinations as *? minutny galdež < gvalt, gomon> 'a minute din <uproar, hubbub>'*. It shoud be said that the following phrases can be realized: *V klasse na minutu podn'als'a galdež <gam, gvalt, gomon>, no učitel' postučal ukazkoy po stolu, I deti pritixli 'There was a hubbub <clamour, uproar, din> in the classroom for a minute, but the teacher tapped the desk and the children got quiet'*. But prototypically all these nouns describe a sound lasting a lengthy period. Compare also: *Na ploščadi den' i noč stoyal gam <galdež, gvalt> 'There was a clamour <hubbub, uproar> on the square day and night'; Za oknom ves' den' ne stixal detskiy <ptičiy> gomon 'The whole day there was a children's <birds'> hubbub heard outside'*.

Two nouns out of the words in questions **GAM 'clamour'** and **GOMON 'hubbub'** denote quite homogeneous noise that can be defined as follows as follows: **GAM** means 'a loud homogeneous prolonged noise that can occur when people speak loudly and simultaneously so causing difficulty in singling out separate words'; **GOMON** is 'homogeneous noise serving as background for other sounds/noises that can occur when many people speak simultaneously thus causing difficulty in singling out separate words'. However, it should be noted, that combinations of these nouns with adjectives denoting complete homogeneity of the sound are impossible as neither **GAM**, nor **GOMON** are "perfectly" homogeneous sounds: **ravnomerny <*monotonny> gam <gomon> 'steady <monotonous> clamour <hubbub>'* (compare, for example, *ravnomerny <monotonny> gul 'steady <monotonous> hum'*). (But here it can be assumed that the impossibility of the combination ** monotonny gomon 'monotonous hubbub'* is determined firstly by the fact that *monotonny* is associated with the notions of 'dull', 'tiresome', 'bothersome', meanwhile *gomon* possesses quite opposite associations. Compare: *vesely <radostny, privetlivy> gomon 'jolly <joyful, affable> hubbub'*, but *?nadoyedlivy <*ugnetayuščiy> gomon 'tiresome <depressing> hubbub'*.

[1] (Lyaševskaya, 2004, p. 315).

The other two words **GALDEŽ 'hubbub', 'din'** and **GVALT 'uproar'** denote heterogeneous noise: *galdež* and particularly *gvalt* can manifest considerable single "bursts" of volume. Compare corresponding parts of the definitions: **galdet'** *'make a hubbub'* **1** (**galdež 1** is a derivative from this verb – compare *galdež gračey 'rooks' clamour'* – and is defined through the verb; there should be a link to the verb in the dictionary entry) means 'to shout/cry simultaneously thus creating heterogeneous noise in terms of volume and pitch' [about birds]; **galdet' 2** (connected with **galdež 2** in the same manner; compare *galdež detvory 'kids' din'*) is 'to speak loudly and simultaneously thus creating heterogeneous noise in terms of volume and pitch'. **Gvalt** is 'a strong unpleasant noise heard in clamour, heterogeneous in terms of volume and pitch heard in clamour' [2].

So for being included into the singularia tantum type, the most important factor for such nouns is neither the homogeneity of the sound, nor the regular repetition of its pitch movement, nor, perhaps, the considerable length of it, but the continuity of its existence. Compare: *bespreryvny <besprestanny> gam < galdež, gvalt>, nesmolkaemy gomon 'endless <non-stop> clamour <din, uproar>; never-ceasing hubbub'*. (However, all three Russian adjectives – *bespreryvny, besprestanny, nesmolkaemy 'endless', 'non-stop', 'never-ceasing'* – express at the same time the idea of length/duration).

The fact that the nouns in question describe sound as indiscrete continuum puts a semantic veto on creating their plural forms. Continuity, thus, presents the most important sound property than the other above mentioned ones.

It should be noted here that all the nouns under consideration suggest only a collective sound producer. Compare: *V barake galdež: u kogo-to payku dnem uveli, na dneval'nyx kričat, I dneval'nye kričat 'In the barrack there's a din: somebody's ration has been pinched, the guys on duty are being shouted at, and the guys on duty are shouting' (A. Solzhenitsyn*); *Kak tol'ko perestupit ona porog, ee vstretit gamom I svistom celaya tolpa 'As soon as she comes out, a crowd will welcome her crying (in a clamour) and whistling' (L. Andreev)*; *Vdrug vse l'udi, vyzvannye na podium, stali čto-to kričat', podn'als'a žutkiy gvalt 'All of a sudden all the people on the stage began shouting something; a terrible uproar arose' (L. Petrushevskaya)*; *V koridore r'adom s kabinetom razdavals'a topot I gomon, kto-to podergal dver', poslyšalis' šlepki brošennyx na pol portfeley 'In the corridor next to the office there was a patter of feet and a hubbub heard, someone tried to pull the door, one could hear slaps of briefcases dropped on the floor' (A.Ivanov)*[3].

The noun **OR 'shouting', 'yelling'** also belongs to the first group. It corresponds to the three meanings of the verb **ORAT' 'shout, 'yell'**: **orat' 1** - *A1 yells from A2* 'Living being A1 shouts loudly and in a drawling manner very often due to a strong emotion or pain A2; the speaker feels unpleasant'; **orat' 2** - *A1 shouts A2 to A3* 'Person A1 pronounces very loudly and in a drawling manner utterance A2 addressing to person A3; the speaker feels unpleasant; **orat' 3** *A1 shouts at A2* 'Person A1, being very displeased with the behavior of A2, tells this

[2] *Ibid.*, pp. 633, 634.

[3] All the examples discussed in the paper and having a reference to the author are taken from the National Russian Language Corpus (www.ruscorpora.ru).

to him/her very loudly and in a specific tone; the speaker appreciates this negatively'[4]. But the derivative noun has not evolved separate meanings corresponding to the stated meanings of the verb.

On the whole, this noun is associated with the idea of a prolonged sound and in this respect it is similar to the above mentioned words. Compare standard uses: *Beskonečny <postoyanny, neskončayemy> or 'endless <constant, never-ceasing> shouting'; Krugom stoyal or 'There was shouting all around'; Ot polučasovogo ora premier vzmok, rubaška prilipla temnymi p'atnami 'Because of a thirty minute shouting the PM got soaked, with his shirt stuck to the body' (A.Voznesensky)*; as well as its figurative use: *Ves' etot or tak I ne prekraščals'a s momenta vyxoda knigi 'All that shouting has been going on since the book was published' ("The Swan")*. On the contrary, the following collocations are impossible: **Razdals'a korotkiy <sekundny, otryvisty> or 'There was a short <a second-long, brief> shouting'*.

However, **OR** is significantly different from **GALDEŽ**, **GAM**, **GVALT** and **GOMON**. **OR**, as a rule, manifests a sequence of very loud and lengthy cries, which of them is pronounced in one breathing out and eventually ends in a pause. Such sequence can take any lengthy period, but, from an objective point of view, it is discrete. Compare: *Vs'u noč iz komnaty donosils'a košačiy or 'All night through there was a cats' yawl heard'*.

On the other hand, **OR 'shouting', 'yelling'** can be perceived as a single loud lengthy cry. But such usage is practically on the confines of the word's semantics. For instance, the statement *Razdals'a košačiy or 'There was a cats' yawl heard'* is quite acceptable, when *?Kot izdal or ?'The cat uttered a yawl'* is doubtful. In the both cases one quantum of the corresponding sound is in the point, but the collocation with the verb *izdat' 'to utter'*, in contrast with *razdat's'a 'to be heard'*, accentuates the single character of the sound (*izdat'* – one sound can be uttered, but *razdat's'a* - either one sound, or a sequence can be heard) – thus the usage of **OR** gets limited.

But, on the whole, **OR** in comparison with **GALDEŽ**, **GAM**, **GVALT** and **GOMON** is, so to speak, a step closer semantically to the nouns **KRIK 'cry', 'shout'** and **VOPL' 'scream', 'yell'**, which mean single separate sounds (as it is to be shown later the noun **KRIK** can also be used to denote not a single sound, but nevertheless its central meaning is a separate sound).

Let us consider the phrase *Posle nedolgogo <minutnogo> ora mladenec zamolk 'After a short/ a minute cry the baby went silent'*. Similar uses are believed to sound more natural than those mentioned above *?minutny galdež <gvalt, gomon> ? 'a minute clamour <din, hubbub>'*. The reason for that is that **OR** with its first meaning of a lengthy sequence of separate sounds, but, at the same time, with a potential of denoting a single separate sound can readily allow contexts assuming short periods of sounding.

However, as well as **GALDEŽ**, **GAM**, **GVALT** and **GOMON**, this lexeme belongs to the singularia tantum type. In spite of its objective discreteness, **OR**, being a sequence of frequently repeated sounds, is reflected in the language as indiscrete sound. Compare standard collocations: *bespreryvny <besprestanny, beskonečny, neskončayemy> or 'non-stop <endless, never-ending> shouting'* – here it is easy to trace that the noun under consideration

[4] *Ibid.*, p. 637.

can be combined with the same adjectives (or similar adjectives) as **GALDEŽ**, **GAM, GVALT** and **GOMON** can.

OR can infer any number of sound producers. Compare, on the one hand, *On podn'al čudoviščny or 'He raised an appalling cry <shouting>'* and, on the other, *Vse-taki ya prorvals'a čerez vseobščiy or i skazal, čto pročitayu stixi 'Finally I got through that shouting all around and said that I was going to recite my poetry' (A Voznesensky)*. This property singles out the word from the other four nouns under consideration belonging to the first group, but anyway this property makes it closer to the nouns **KRIK 'cry', 'shout'** and **VOPL' 'scream, 'yell'.** It is quite obvious that a potential of denoting a single sound producer and a potential of denoting a single sound are interrelated: to produce a single separate sound is natural for a single sound producer, and if there are several subjects, then special efforts to synchronize are necessary for producing such sound.

Let us consider now properties of the nouns **KRIK** and **VOPL'**. Shall we start with the noun **KRIK**.

This word can be used in its three meanings. Firstly, it can denote a sound pronounced by a person or an animal with a strong breathing effort; as a rule, such sound is produced because of pain or a strong emotion. Compare: *krik radosti <vostorga, boli, gneva, otčayaniya> 'cry of joy <delight, pain, anger, desperation>'*. Secondly, **KRIK** can describe an utterance pronounced by a shouting person and which is addressed to another person. Compare: *Razdals'a krik: "Ey, podoždite!" 'There was a cry: "Hey! Wait!" heard'*. And, finally, in its third meaning the noun **KRIK** can denote a person's displeasure by some situation and which is pronounced by a loud voice and in a special tone with its aim to improve the situation. Compare: *Esli ya pridu domoy pozdno, budet krik 'If I come home late, there will be a lot of shouting'*[5].

Let us consider this word in terms of its properties of duration and discreteness / indiscreteness.

In its first and second meanings the noun **KRIK** can equally describe both a long and a short sound. Compare: *otryvisty <korotkiy> krik – protyažny krik 'A staccato <short> cry' – 'a long-drawn-out cry'* [in both cases it can be **krik 1** or **krik 2**]; *Nikto ne pomnil, skol'ko dlils'a etot dušerazdirayuščiy krik 'Nobody could recall for how long that harrowing cry could be heard'* [**krik 1**]; *Nad tribunami povis protyažny krik: "Go-o-ol!" 'There was a long cry heard over the stadium: "Goal!"* [**krik 2**]. It should be noted here, that the duration of *krik* in similar uses is basically different from the duration expressed by the above mentioned words. *Krik* here presents a sound formed by a single breathing out and this fact considerably limits its duration. In other words, in all the adduced examples the noun *крик* indicates "a quantum" of the sound. Compare: *pervy krik rebenka 'the baby's first cry'* [**krik 1**]; *Razdals'a krik: "Stoy!" 'There was a cry: "Freeze!"* [**krik 2**].

But there exists another usage of the word in question – when **KRIK** describes multiple repeated sounds or a range of various utterances and in this way gives an idea about a

[5] *Ibid*., p. 635.

multitude of separate quanta; either meaning of **KRIK** can be realized in such uses. Moreover, the third meaning of **KRIK** can be realized only in this usage. Compare: *Mladenec kričal vs'u noč, i yego krik ne daval nikomu spat' 'The baby cried all night through and his cries let nobody fall asleep'* [**krik 1**]; *So vsex storon donosils'a krik torgovok 'There cries of market sellers all around'; V dome podnyals'a detskiy krik: "Ura! My edem na more!" 'In the house there were the children's cries heard: "Hurrah! We are going to the sea!"* [in both cases it is **krik 2**]; *Ix vstreči vsegda končalis' krikom ili drakoy 'All their meetings would end up in shouting and fighting'* [**krik 3**].

Thus, the noun **KRIK**, as well as **OR,** can denote a discrete sequence or a range of sounds or utterances, but the language reflects it as indiscrete continuum (the same happens with **OR**). In this usage both **KRIK** and **OR** do not possess the plural form.

As well as **OR**, the noun **KRIK** in all its meanings can infer one or many sound producers. Compare, for instance, *krik žertvy – krik žertv 'a victim's cry' – 'the victims' cry''*. The latter case (when the singular form is used in a distributional context) can present two possible interpretations: firstly, it can be single (simultaneous) common shouting; secondly, it can be a sequence of *cries*-quanta that are produced by various subjects in different times but they merge into one common, so to speak, endless shouting.

Now let us consider what the plural form of this noun means.

The plural form indicates a multitude of single *cries* and, as well as the singular form, can correspond to any number of sound producers. Thus, the plural form of this noun can be synonymous to its singular form in case the latter denotes a discrete sequence (it should be noted once more, that this sequence is objectively discrete) of sounds or utterances. Compare, on the one hand, the usage of the singular form: *Vs'u noč pod oknom razdavals'a čey-to krik 'All night through there was someone's cry heard outside'* [**krik 1** or **krik 2**]; *Vse utro v dome stoyal krik: "Gde moi očki? Opyat' ničego ne mogu nayti!" 'All the morning there was shouting heard in the house: "Where are my spectacles? I can't find anything here again!"* [**krik 2**]; and, on the other hand, the usage of the plural form: *V otvet razdalis' kriki: "Nam ne nužno muzyki, nam nužen xleb, den'gi I novye doma" 'There were cries heard in reply: "We don't need music, we need bread, and money, and new houses" (S.Spivakova))* [**krik 2**; in this case the plural form indicates multiple *cries* produced by a multitude of sound producers; the direct speech conveys general sense of the ideas expressed by all the present]. But **krik 3** does not possess such pair as in this meaning **KRIK** (as we have already noted) does not have the plural form. Compare: *'Čto ty skandališ? Nadoel tvoy postoyanny krik!' 'Why should you make fuss again? I am sick and tired of your never-ending yelling!'*

A slight semantic difference between the forms of singular and plural consists in the point that the singular form indicating a multitude of single short acts, nevertheless, presents them as a persistent continuum, as one prolonged *cry*; when the plural form clearly conveys the idea of discreteness.

As it was stated by E. Rakhilina[6], in pairs similar to the type *krik – kriki 'cry' – 'cries'* the plural form is used to convey a typically aspectual meaning – the meaning of iterativity (in the

[6] (Rakhilina, 2000, p. 67).

paper the example of *vzdox - vzdoxi (sigh – sighs)* is drawn that is identical to the pair *krik – kriki (cry – cries)* related to the topic of interest). O. Lyashevskaya introduces the term "the iterative plural form" for the case in question and states that such usage of the plural form is not connected with the category of countability /uncountability[7].

The noun **VOPL' 'scream, 'yell'** which is a syntactical derivative of the verb **VOPIT'** 'yell or scream often because of a strong emotion or pain' [8], in contrast to **KRIK**, can denote only a single sound or a quantum. Compare a standard usage: *On izdal vopl' 'He made a scream'* and an impossible usage: **Vs'u noč pod oknami slyšals'a čey-to vopl' *'Outside there was a scream heard all night through'*.

Scream/yell can last either a long or a short period, but even for a long *scream/yell* there is a similar limit as to the quantum of *cry*: its duration is limited by the fact that this sound is pronounced in one breathing out. Compare: *protyažny vopl' 'A long-drawn-out scream/yell'; Razdals'a dlinny nizkiy vopl', opuskayuščiys'a do utrobnogo ryčaniya 'There was a long low scream heard that was going into deep and hollow growl' (L.Ulitskaya)*; and, on the other hand, *otryvisty <korotkiy> vopl' 'a brusque/short scream'*.

As well as **KRIK**, the noun **VOPL'** can infer one or many sound producers. Compare: *vopl' žertvy – vopl' žertv ' victim's scream' – 'victims' scream'*. However, in comparison with the collocation *krik žertv 'victims' cry'* the latter case can be interpreted only in one way: there is only one synchronous common *scream/yell*, but not a sequence of this sound's quanta, as the noun **VOPL'** can indicate only a single quantum.

The plural form of *scream/yell* indicates a multitude of single sounds and also infers an optional number of sound producers. As in its singular form the noun **VOPL'** (in comparison with **KRIK**) cannot denote multiple sounds, so the plural form in no case synonymous to the singular form. Compare: *vopl' bolel'ščikov 'the fans' yell'* [only single sound] – *vopli bolel'ščikov 'the fans' yells'* [only multiple sound].

So we have considered the following semantic properties of nouns denoting voice sounds: duration of its existence; homogeneity/heterogeneity of the sound (for those nouns with this property being actual); discreteness/indiscreteness; the number of sound producers involved. Besides that, we have considered one grammatical characteristic of the nouns in question – their potential of creating the plural form.

The nouns under consideration are heterogeneous in terms of the properties stated above. One side is presented by the nouns **GALDEŽ 'hubbub', din'**, **GAM 'clamour'**, **GVALT 'uproar'**, **GOMON 'hubbub'**. All these nouns indicate a lengthy sound created by a multitude of sound producers. This sound is perceived as either quite homogeneous, though not ideally, (**GAM**, **GOMON**) or having considerable volume fluctuations (**GALDEŽ**, **GVALT**), but, in any case, this sound is conceptualized as indiscrete continuum. The above mentioned nouns do not allow creating plural forms.

[7] (Lyashevskaya, 2004, p. 313).

[8] (Prospekt aktivnogo slovarya, 2010, p. 632).

The other side is presented by the noun **VOPL' (scream/yell)**. This noun indicates sounding of any duration including short staccato sounds, infers any number of sound producers and describes only one quantum of the sound produced in one breathing out. That is why even a lengthy *scream/yell* according to the, so to speak, "ultimate scale" is considerably shorter those sounds that are denoted by the four other nouns mentioned above. This noun possesses the both forms.

From the rest of the words the noun **OR (shouting)** is close to the first side and the noun **KRIK (cry, shout)** – to the other.

OR in its standard uses denotes a lengthy sound created by either a multitude of sound producers or by only one sound producer and consists of a sequence of sound quanta. A possibility to denote a single quantum is practically out of the semantic potential of this word, but in contrast to the first side nouns is sometimes possible. As well as these nouns **OR** conceptualizes a sequence of sounds as indiscrete one and does not possess the plural form.

KRIK, as well as **VOPL'**, indicates a sound of any duration and infers any number of sound producers but it can describe not only a single quantum, but an indiscrete continuum as well. In the former case the noun **KRIK** possesses the both forms of number, but in the latter it does not have the plural form.

To conclude, those nouns that describe a speech situation as a process but not as a set of single completed acts do not possess plural forms.

Acknowledgements

This research has been financed by a research program of History and Philology Branch of the Russian Academy of Sciences, a grant from the Russian Humanitarian Scientific Foundation (No. A 10-04-00273), and by the Russian President grant for the Support of Leading Scientific Schools No. НШ-4019.2010.6.

Bibliography

Lyaševskaya, O. N. 2004. *Semantika russkogo čisla*. Moscow.

Rakhilina, E. V. 2000, *Kognitivny analiz predmetnyx imen: semantika I sočetayemost'*. Moscow.

Prospekt aktivnogo slovarya russkogo yazyka pod red. Ju. D. Apresyana. 2010. Moscow.

Interpretative Verbs and Interpretative Constructions with Converb Clauses

Tilmann Reuther

Slavic Department, Alpen-Adria-Universität Klagenfurt, Austria
tilmann.reuther@uni-klu.ac.at

Abstract

The paper deals with interpretative verbs as established by (Apresjan 2004) and interpretative converb constructions as established by (Boguslavskij 1977) and afterwards discussed in a cross-linguistic typological perspective in (Haspelmath & König 1995). It is shown that Apresjan's approach offers a key to the semantics of converb (DEEPR) constructions. Special attention is paid to converb constructions of the V – DEEPR type with postponed DEEPR clause and both V and DEEPR in the perfective verbal aspect (of the type *On prosčitalsja, poexav na avtobuse* 'He made a mistake, having gone by bus'), and their syntactic equivalents.

Keywords

Semantic verb classes, converb constructions, verbal aspect.

1 INTRODUCTION

1.1 Interpretative Verbs

Working with a fundamental classification of predicates (cf. Apresjan 2003, 2006) Ju. D. Apresjan established the class of interpretative verbs as one of the main verbal classes of almost the same rank as verbs with the meaning of action (dejstvie), activity (dejatel'nost'), behaviour (povedenie), occupation (zanjatie), impact (vozdejstvie), process (process), manifestation (projavlenie), position in space (položenie v prostranstve), state (sostojanie), quality (svojstvo), parameter (parametr), existence (suščestvovanie) etc. (cf. Apresjan 2004:8). The lexicographic definition of an interpretative verb has a standard form with one part – the presupposition P, and a second part – the assertion R. Let us look at an example of such a definition (Apresjan 2004:9):

(1) *X pooščrjaet Y-a, delaja P* = 'X sdelal P [presuppozicija]; govorjaščij sčitaet, čto P otnositsja k klassu dejstvij, pokazyvajuščix, čto čelovek, kotoryj ix soveršaet, odobrjaet dejstvija ili dejatel'nost' drugogo čeloveka i xočet pobudit' ego prodolžat' dejstvovat' tak že [assercija]'

> *X encourages Y, doing P* = 'X did P [presupposition]; the speaker thinks that P belongs to the class of actions which show that a person who completes them wellcomes the actions or activities of another person and wants to stimulate this person to continue doing so [assertion]' (Translation – *T.R.)*

As one can see the so called 'interpretation' is introduced by the component 'the speaker thinks'. In the following examples from the National Corpus of Russian (NCR)[1]

(2) *Bolee togo, gosudarstvo ėtu dejatel'nost' pooščrjalo, osvobodiv „star'evščikov" ot naloga.* /Evgenij Borisenkov. Metalloiskateli (2004) // „Za rulem", 2004.03.15

'Moreover, the state encouraged this activity, having exempted the „ragmen" from tax.'

we have X = *gosudarstvo* 'state', P = *osvobodit' ot naloga* 'exempt from tax', Y = *dejatel'nost'* 'activity'[2],

(3) *Osobenno sblizilis' oni s Dilejny, i tot pooščrjal Erika bol'še pet', bol'še pisat' pesni i v konce koncov polučat' kajf ot togo, čto on muzykant.* /Cena ljubit gitarista (2002) // „Drugoj", 2002.11.15

'They especially chumed up with Delany, and he (=Delany) encouraged Erik more to sing, more to write songs, and in the end to get satisfaction from the fact that he (is) a musician.'

we have X = *Dilejny* 'Delany', P is not stated explicitly[3], Y = *Erik.*

Going into lexical semantics (Apresjan 2004:11) distinguishes among several types of interpretation:

a) ethic interpretation (the most numerous group): *pomogat'* 'to help', … , *pokrovitel'stvovat'* 'to patronize'; … ; *podvodit' (kogo-l.)* 'to betray (somebody)', … ; … ; *balovat' (rebenka)* 'to spoil (a child)'; … ; *oskorbljat'* 'to offend', … ; *izdevat'sja* 'to mock', … ; … ; *nakazyvat'* 'to punish', … , *pooščrjat'* 'to encourage'; *zloupotrebljat' (doveriem)* 'to abuse (somebody's confidence)';
b) juridical and religious interpretation: *narušat' pravila* 'to infringe the rules', … ; … ; *grešit'* 'to commit a sin', …; … , *soblaznjat'* 'to seduce';
c) logical, or truth-conditional interpretation: *ošibat'sja* 'to make a mistake', … ; *preuveličivat'* 'to exaggerate', … ; *nedoocenivat'* 'to underestimate', *pereocenivat'* 'to overestimate';
d) utilitaristic interpretation: *vyigryvat'* 'to win', … ; *(po)gorjačit'sja* 'to get excited, to overreact', … ; *oplošat'* 'to misjudge', … , *promaxnut'sja* 'to fail to hit the goal';

---

[1] All examples from the NCR www.ruscorpora.ru were taken on June 10, 2011.

[2] Here, the actant Y is an abstract noun, not a person. According to the data from NCR this kind of construction is much more frequent than the construction with names of persons. However, it is clear that the activity is assigned to the persons called "*star'evščiki"*, cf. *Bolee togo, gosudarstvo pooščrjalo „star'evščikov", osvobodiv ix ot naloga.* 'Moreover, the state encouraged the „ragmen", having exempted them from tax. '

[3] As stated by (Apresjan 2004:9) this is a quite regular situation (the actant P being implied by the context).

e) combined interpretation (mostly a combination of ethic and logical interpretation): *izobražat' v černom cvete* 'to depict in black color', … ; *priukrašivat'* 'to prettify', … ; *obmanivat'* 'to deceive', … , *krivit' dušoj* 'to dissemble one's feelings'.

Apresjan investigates aspectual properties of (prototypical) interpretative verbs. Their most important aspectual characteristic is perfectivity (perfektivnost'), i.e. when used in the form of NESOV NAST (imperfective aspect, present tense) with reference to the moment of speech, most interpretative verbs convey the perfective meaning (perfektnoe značenie), and not the actual-durative one: *Vy ošibaetes' <predaete obščie interesy, postupaete nizko>* 'You are making a mistake <betraying common interests, acting meanly'> means that the person has already done something which is interpreted as a mistake, a betrayal of common interests, or meanness (Apresjan 2004:6, 17f.).

Further on, (Apresjan 2004:18f.) discusses several syntactic characteristics of interpretative verbs. Most importantly, the valency P, if expressed explicitly, comes in five different ways: Either a) as a converb construction (*On prosčitalsja, poexav na avtobuse* 'He made a mistake, having gone by bus'; *Vy preuveličivaete, govorja, čto p'esa provalilas'* 'You are exaggerating, saying that the play fell through'), or b) as a subordinate clause with the conjunctions *esli* 'if' or *kogda* 'when' (*Vy preuveličivaete, kodga govorite, čto p'esa provalilas'* 'You are exaggerating when you say that the play fell through'), or c) as a coordinative chain (*Devica mešala emu vesti mašinu – bez umolku taratorila, vertelas', xvatala za ruku* 'The girl disturbed him to drive the car – she unceasingly jabbered, hovered around, grabbed his hand'), or d) as a pseudocoordinative chain of the type *P i tem samym R* 'P and thereby R' (*On opozdal i tem samym vsex podvel* 'He came late and thereby let us/them all down'), or e) as a colloquial construction with an anaphoric sententional pronoun of the type *ėto* 'that', *tut* 'here' (*Ėto ty pogorjačilsja* 'That you overreacted'*; Tut ty oplošal* 'Here you misjudged').

Type a), i.e. the converb construction, brings us directly to the following Section 1.2.

1.2 Converb Constructions

In Russian, constructions with a finite verb (V) and a converb (also called adverbial participle, in Russian deepričastie - DEEPR) can come in 2 x 4 x 2 = 16 different sentence types according to the following scheme:

DEEPR: **verbal aspect**	**V:** **verbal aspect and tense**	**Position of DEEPR clause** **relative to V**
	SOV	
SOV	NESOV	PREPOS
NESOV	PROSH	POSTPOS
	NEPROSH	
2	4	2

where SOV – perfective verbal aspect, NESOV – imperfective verbal aspect; PROSH – past tense, NEPROSH – non-past tense, i.e. present or future tense; PREPOS – DEEPR clause precedes V, POSTPOS – DEEPR clause follows V.

For the purpose of this paper we will have a look at four types from the above scheme – constructions with both the finite verb and the converb in the perfective verbal aspect, and the converb clause either preceding, or following the main clause. Let us begin with two preposed and one postposed converb clause:

(4) V_SOV – DEEPR_SOV/PROSH_PREPOS
Otorvavshis' ot bumag, on vzgljanul na Efimovu. (NCR)

'Having turned away from the papers, he looked at Efimova.'

(5) V_SOV – DEEPR_SOV/NEPROSH_PREPOS
… porjadochnaja zhenshchina, razgljadev duraka, perestanet im zanimat'sja. (Akimova & Kozinceva 1987:273)[4]

'… a decent woman, having made out a fool, will give up associating with him.'

(6) V_SOV – DEEPR_SOV/PROSH_POSTPOS
Efimova vyshla, ne poproshchavshis'. (NCR)

'Efimova walked out, not having said „Good bye".'

Looking at the iconic-chronological „figure"[5] of sentences (4) - (6) I agree with the point made by (Rappaport 1984:90):

(7) "There is a natural iconic relation between linear order, on the one hand, and temporal or teleological order, on the other. Linear anteriority can be associated with temporal anteriority, and linear posteriority – with temporal posteriority. Similarly, since a means is logically prior to its consequence, linear anteriority can be associated with a means, and linear posteriority – with its consequence. These iconic relations can be violated when the AvPrt (adverbial participle, i.e. converb – *T.R.)* clause is postposed, but not when it is preposed. Thus, in the relevant aspects, an initial AvPrt clause must observe iconicity, while a final AvPrt clause need not do so."

Let us now have a closer look on the case of the final AvPrt clause, i.e. the postponed DEEPR clause, and link our considerations to an example discussed in (Boguslavskij (1977:271). To my knowledge, I.M. Boguslavskij was the first to define the interpretative meaning for converb constructions, his example being the following:

(8) *On sygral na ruku pravym, perenesja <tem, chto perenes> srok obsuzhdenija zakonoproekta.*

'He played into the hands of the right-wingers, having moved <by the fact that he moved> the date of the reading of the bill draft.'

With respect to the semantics of this sentence, (Boguslavskij 1977:271) states:

---

[4] As one can see from this example, inserted converb clauses are classified according to their position relative to the verb in the main clause.

[5] I use the term „figure" in order to refrain from terminological debates on tempus and taxis.

(9) „ … there is only one event (sobytie) A which is interpreted (interpretirujetsja) by the speaker as B. In other words, B consists (or manifests itself) (zakljuchaetsja (ili projavljaetsja) in A.“ (Translation from the Russian original – *T.R.)*

In other words, the postponed converb construction (8) with interpretative meaning gives us an obvious example of an iconic-chronological „figure“ where linear posteriority does not mean temporal posteriority. But what about linear anteriority?

Most obviously, the inversion of the main and the converb clauses of sentence (8) delivers a perfect synonymous paraphrase, cf.

(10) *Perenesja <Tem, chto perenes> srok obsuzhdenija zakonoproekta, on sygral na ruku pravym.*

‘Having moved <By the fact that he moved> the date of the reading of the bill draft, he played into the hands of the right-wingers.’

So what about the part of Rappaport’s rule (7) that an initial converb clause must observe iconicity? The solution to this question is the fact that the phraseme *ИГРАТЬ НА РУКУ* ‘TO PLAY INTO THE HANDS’ belongs to the class of interpretative predicates, its sententional form being the following: *X играет на руку Y-у, делая P* ‘X plays into the hands of Y, doing P’. In the above examples (8) and (10) X = *on* ‘he’, Y = *pravye* ‘the right-wingers’, P = *perenesti srok obsuždenija zakonoproekta* ‘to move the date of the reading of the bill draft’, and – according to Apresjan’s scheme – P is the presuppositional part of the lexicographic definition of the ***single*** situation described by the interpretative phraseme in question.

As a consequence, the iconic-chronological „figure“ of converb constructions must be described in a more detailed way. I will try to do this by discussing the case of „interpretative“ converb constructions.

1.3 Interpretative Verbs and Interpretative Converb Constructions

1.3.1 Interpretative converb constructions with interpretative verbs

According to Apresjan (see Section 1.1 above) the expression of the valency P of interpretative verbs in the form of a converb clause is one of the regular cases. In other words, the ‘interpretative’ semantics of a converb clause that depends on an interpretative verb is based on its ***actant status*** in relation to the predicate of the main clause. The iconic-chronological “figure” of the complex sentence is one ***single*** situation, and it is only the ***internal*** chronological ordering of the components of the interpretative verbal meaning which can be applied. Let us remember that in Apresjan’s definition of the verb *pooščrjat’* ‘to encourage’ the doing of P only „internally“ precedes the interpretation proper.[6]

[6] To my opinion, ‘X sdelal P [presuppozicija] ‘X did P [presupposition]‘ is not the only proper way to define the presupposed event P. It seems closer to the truth to allow for the following alternative: ‘X sdelal <načal delat’, delaet> P [presuppozicija] ‘X did <began to do, does> P [presupposition]‘.

1.3.2 Interpretative converb constructions with non-interpretative verbs?

It seems the case that interpretative converb constructions can also be found with non-interpretative verbs. Consider the following examples from NCR:

(11) *V 1890 godu inzhenery soedinili bachok s siden'em v edinuju konstrukciju, sozdav* ***tem samym*** *proobraz sovremennogo unitaza.*

'In 1890 engineers conjoined the bowl with the seat to a joint construction, having created thereby the prototype of the modern toilet bowl.'

(12) *V nojabre japonskie vojska pererezali Kitajsko-Vostochnuju zheleznuju dorogu (KVZhD), vyzvav* ***tem samym*** *obmen zhestkimi notami mezhdu SSSR i Japoniej.*

'In November the Japanese troops cut the Chinese-Eastern Railway (KVZhD), having caused thereby an exchange of harsh diplomatic notes between the USSR and Japan.'

(13) *Naprimer, v nejtral'nyx vodax mozhno postroit' takoe sooruzhenie, oboznachiv* ***tem samym*** *svoe prisutstvie, pritom nikak ne narushaja normy mezhdunarodnogo prava.*

'For example, in neutral waters it is possible to build such a construction, having marked thereby one's presence, by that in no way infringing the norms of international law.'

The structure of these sentences is the same as in Boguslavskij's example

(14) *On perenes srok obsuzhdenija zakonoproekta, sygrav* ***tem samym*** *na ruku pravym.*

'He moved the date oft he reading of the bill draft, having **thereby** played into the hands of the right-wingers.'

However, it seems clear that neither *sozdat' (proobraz)* 'to create (a prototype)', nor *vyzvat' (obmen)* 'to cause (an exchange)', nor *oboznačit' (prisutstvie)* 'to mark (the presence)' should be called interpretation verbs. Nevertheless, a sort of ***unity*** of the complex situation as expressed by *tem samym* 'thereby' is quite obvious. Cf. also type d) from Apresjan's syntactic list – the pseudocoordinative chain of the type *P i tem samym R* 'P and thereby R' (*On opozdal i tem samym vsex podvel* 'He came late and thereby let us/them all down').

As a consequence, there is one question to be solved: Why do non-interpretative verbs like *sozdat' (proobraz)* 'to create (a prototype)', *vyzvat' (obmen)* 'to cause (an exchange)', and *oboznačit' (prisutstvie)* 'to mark (the presence)' easily allow for the interpretative reading of converb constructions? The answer seems to be the following:

Sozdat' 'to create', as used here, has the following actant structure: *X creates Y out of Z for the purpose W; vyzvat'* 'to cause', as used here, has the following actant structure: *X causes Y by Z; oboznačit'* 'to mark', as used here, has the following actant structure: *X marks Y by Z.* In all three cases, the DEEPR clauses in sentences (11) – (13) instantiate the ***actant*** Z, so the *single-situational* reading is easily at hand. I propose to call this unity of situation ***supported*** by the semantics of the connector *P, i tem samym Q* 'P, thereby Q'.[7]

[7] For discussion of single complex situations expressed by two predicates in various syntactic configurations cf. (Poljanskij 1987:250-253; Bondarko 1987; Akimova & Kozinceva 1987: 265-267; Weiss 1993, 1994).

2 INTERPRETATIVE CONSTRUCTIONS AND THEIR SYNTACTIC VARIATION

2.1 Boguslavskij's Rule: The Case of Russian

The phraseme *TEM SAMYM* 'THEREBY' follows Boguslavskij's rule for explanatory words (Boguslavskij 1977:227):

> (15) „In sentences with converb constructions which are in the relation of synonymous paraphrasing, there can be used one and the same explanatory words, in that (prichem) they are attached to one and the same verb, occuring in one case in the final verbal form, and in the other in the converb form." (Translation from the Russian original – *T.R.)*

Cf. from above (14) *On perenes srok obsuzhdenija zakonoproekta, sygrav* ***tem samym*** *na ruku pravym.* 'He moved the date oft he reading of the bill draft, having **thereby** played into the hands of the right-wingers.' equals

> (16) *Perenesja srok obsuzhdenija zakonoproekta, on sygral* ***tem samym*** *na ruku pravym.*
>
> 'Having moved the date of the reading of the bill draft, he played **thereby** into the hands of the right-wingers.'

As a consequence, we can ***add*** the construction of type (14) as a ***sixth*** possible syntactic construction for interpretative verbs – here, the interpretative verb constitutes the postposed DEEPR clause, while the presupposed event P constitutes the preposed matrix clause.

To complete the list of syntactic variation, one instantiation of example (8) above, i.e.

> (17) *On sygral na ruku pravym* ***tem, chto*** *perenes srok obsuzhdenija zakonoproekta.*
>
> 'He played into the hands of the right-wingers by the fact that he moved the date of the reading of the bill draft.'

shall be considered a ***seventh*** possible type of syntactic construction of interpretative verbs – here, the presupposed event P comes as a subordinated clause, linked to the main clause by the two-part conjunction *R tem, chto P* 'R by the fact that P'.

2.2 Interpretative Constructions in German: The Case of a Non-Converb Language Type

Example (17) from above is very close to what is one of the ways to convey the interpretative meaning in German. There is a two-part conjunction – *DADURCH, DASS* 'BY THIS THAT' which serves as a means for connecting the main clause which contains the interpretative predicate and the subordinated clause which expresses the presupposed event P, cf. the German word-by-word equivalent of (17):

> (18) *Er spielte den Rechten* ***dadurch*** *in die Hände,* ***dass*** *er den Termin der Beratung des Gesetzesentwurfs verschob.*

Another connector of less „instrumental“ descendence serves as the ***main*** means to connect the interpretative predicate within the main clause and the subordinated clause which expresses the presupposed event P – *INDEM* ‘IN THAT’, cf. the German equivalent of (17) and (18):

(19) *Er spielte den Rechten in die Hände,* ***indem*** *er den Termin der Beratung des Gesetzesentwurfs verschob.*

For non-interpretative verbs the German conjuction *indem* works the same way, cf., e.g., sentence (11) from above and its Russian and German paraphrases

(20) *V 1890 godu inzhenery sozdali proobraz sovremennogo unitaza, soediniv bachok s siden’em v edinuju konstrukciju.*

Im Jahr 1890 schufen Ingenieure den Prototypen der modernen Toilette, ***indem*** *sie die Schüssel mit dem Sitz zu einer Gesamtkonstruktion verbanden.*

‘In 1890 engineers created the prototype of the modern toilet bowl in that they conjoined the bowl with the seat to a joint construction.’

Another German connector – *WODURCH* ‘BY WHICH’ – is available when it comes to the inverted distribution of the interpretative predicate (now in the subordinated clause) and the event P (now in the main clause), cf. the equivalent of (14)

(21) *Er verschob den Termin der Beratung des Gesetzesentwurfs,* ***wodurch*** *er den Rechten in die Hände spielte.*

‘He moved the date of the reading of the bill draft by which he played into the hands of the right-wingers.’

3 IMPLICATIONS FOR MTT: LEXICON AND GRAMMAR

3.1 Lexicon: Russian and German

In his study on interpretative verbs Ju.D. Apresjan draws a borderline between interpretative verbs, evaluative verbs (ocenočnye glagoly) and verbs of behaviour (glagoly povedenija). Cf. for the following properties (Apresjan 2004:11-14):

- The main difference between interpretative and evaluative verbs is that the two components – an action P and its interpretation R / its evaluation E play different roles in the lexicographic definition: interpretative verbs take P as presupposition and R as assertion, while evaluative verbs take P as assertion and E as modal frame, e.g. (Apresjan 2004:12):

  (22) *To huddle (Jutit’sja)* = ‘to live in a premise, where there is less room than is necessary for normal life [assertion];the speaker poorly assesses the conditions in which the subject is forced to live, or wants the addressee to assess them in this way [modal frame]’ Cf. *The town Grozny was shelled, people huddled in underground stories, without water and light* („Itogi“, 27.08.96) (Translation from the Russian original – *T.R.*).

- Nevertheless, there are verbs which combine both properties, i.e. the above distinction is true only for prototypical cases.

- The lexicographic definition of verbs of behavior like *bezobrazničat'* 'to behave in an improper manner', *bujanit'* 'to raise the roof', *gerojstvovat'* 'to play the heroe', *debošírovat'* 'to paint the town red' falls apart into assertion and modal frame, P forming the assertion, and an interpretation of P making part of the modal frame, e.g. (Apresjan 2004: 14):

  (23) *X hooligans (X xuliganit)* = 'X performs different actions P which disturb the normal existence of other people or are dangerous for them, although do not endanger their life [assertion]; the speaker thinks that P heavily infringes the norms of social behaviour and that X behaves in this way on purpose; therefore the speaker assesses the behaviour of X harshly [modal frame]' Cf. *They hooliganed in the streets, offended passers-by, performed different wild fooleries, and in general were not able to behave properly*. (N. Nosov) (Translation from the Russian original – *T.R.)*.

All these observations on ***lexical*** semantics can be applied to both Russian and German, and to English, too.

- When it comes to the semantics of ***grammatical*** categories, Russian aspect plays a crucial rule, and such properties cannot apply to typologically different verbal systems like those of German or English. Cf. on Russian (Apresjan 2004:14):

  „Every behaviour presupposes the observability (nabljudaemost') of what a person really does, in that (prichem) one usually speaks about a behaviour when one sees a series of single-type acts (rjad odnotipnyx aktov) of a person or another living being over the period of one round of observation (na protjazhenii odnogo raunda nabljudenija); cf. *to balk (artachit'sja), to paint the town red (deboshirit'), to buffoon (pajasnichat')*. Therefore, behaviours, in contrast to interpretative and most of evaluative verbs can freely be used in the actual-durative meaning of the IMPERFECTIVE aspect. Cf. *Look how she is grimacing <is behaving capriciously> (Posmotri, kak ona krivljaetsja <kapriznichaet>), Stop grimacing <behaving capriciously> (Perestan' krivljat'sja <kapriznichat'>), When the police came the crowd was still roistering (Kogda pribyla policija, tolpa vse eshche beschinstvovala)* etc." (Translation from the Russian original – *T.R.*).

3.2 Grammar: Russian and German

Most obviously, the ways to convey the meaning of 'interpretation' in Russian by converb constructions, and the need to use different connectors in German put a sensible challenge to grammarians, above all for those working on systems of automatic translation under the Meaning-Text-approach. This paper, being devoted to the lexicon, is not the place to elaborate on this point.

4 CONCLUSION

We were able to show that the meaning of 'interpretation' is important for both the lexicon and the grammar and that the lexicographic definition of Ju. D Apresjan as presented in (Apresjan 2004) is a key to the understanding of Russian converb constructions and their syntactic equivalents in Russian and German.

Acknowledgements

I owe my special thanks to the research team of the Linguistic Laboratory of the Moscow-based Institut problem peredači informacii (IPPI) for the possibility to work on the ETAP-3 machine and to give a talk on DEEP constructions in April 2010. A previous version of this paper was reviewed by Valentina Apresjan, Igor' Mel'chuk and Daniel Weiss; their comments and criticism helped me to improve my presentation and correct some points in it.

Bibliography

Akimova, T. G. & N. A. Kozinceva 1987. Zavisimyj taksis (na materiale deepričastnyx konstrukcij). In *Teorija funkcional'noj grammatiki. Vvedenie, aspektual'nost, vremennaja lokalizovannost', taksis*, 257-274. Leningrad: Nauka.

Apresjan, Ju. D. 2003. Fundamental'naja klassifikacija predikatov i sistemnaja leksikografija. In *Grammatičeskie kategorii: ierarxii svjazi, vzaimodejstvije. Materialy meždunar. nauč. konferencii*, 7-21, Sankt-Peterburg. Revised and enlarged version in Apresjan 2006:75ff.

Apresjan, Ju. D. 2004. Interpretacionnye glagoly: semantičeskaja struktura i svojstva. *Russkij jazyk v naučnom osveščenii*, 1(7):5-22. Revised version in Apresjan 2006: 145ff.

Apresjan, Ju. D. (ed.) 2006. *Jazykovaja kartina mira i sistemnaja leksikografija*, Moskva: Jazyki slavjanskix kul'tur.

Boguslavskij, I. M. 1977. O semantičeskom opisanii russkix deepričastij: neopredelennost' ili mnogoznačnost'? *Izvestija AN SSSR, Ser. lit. i jaz.*, 36/3:270-281.

Bondarko, A.V. 1987. Zamechanija ob otnoshenijax nedifferencirovannogo tipa. In *Teorija funkcional'noj grammatiki. Vvedenie, aspektual'nost, vremennaja lokalizovannost', taksis*, 253-256. Leningrad: Nauka.

Haspelmath, M. & E. König. 1995. *Converbs in Cross-Linguistic Perspective. Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerund*, Berlin – New York: Mouton de Gruyter.

National Corpus of Russian = Nacional'nyj korpus russkogo jazyka www.ruscorpora.ru

Poljanskij, S. M. 1987. Odnovremennost'/raznovremennost' i drugie tipy taksisnyx otnoshenij. In *Teorija funkcional'noj grammatiki. Vvedenie, aspektual'nost, vremennaja lokalizovannost', taksis*, 243-253. Leningrad: Nauka.

Rappaport, G.C. 1984. *Grammatical Function and Syntactic Structure: The Adverbial Participle of Russian.* Columbus (Ohio): Slavica

Weiss, D. 1993. Aus zwei mach eins. Polyprädikative Strukturen zum Ausdruck eines einzigen Sachverhalts im modernen Russischen. In Ebert, K. (ed.), *Studies in Clause Linkage. Papers from the First Köln-Zürich Workshop*, 219-238, Zürich.

Weiss, D. 1994. Die Vielfalt der Einheit (zwei Konjunkte, ein Sachverhalt). In Mehlig, H.R. (ed.) *Slavistische Linguistik 1993*, 307-330, München: Sagner.

A Graph Visualization Tool for Terminology Discovery and Assessment

Benoît Robichaud

Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal

benoit.robichaud@umontreal.ca

Abstract

This paper presents a Graphical User Interface (GUI) mainly based on a graph visualization device and used for exploring and assessing lexical data found in the DiCoInfo, a specialized e-dictionary of computing and the Internet. Computer visualization devices have been used to present and browse data in many fields, but GUIs for electronic dictionaries have not evolved much. Very few take advantage of the fundamental nature of dictionaries: they are huge and ordered collections of lexical relationships (i.e. *lexical networks*). Graph visualization devices such as intertwined (directed) graphs present themselves as better tools to browse these relationships. They surely are well suited for assessing the consistency of encoded data.

Keywords

Lexical relations, e-dictionary, data visualization, graph model, assessment tool.

1 Introduction*

Electronic support to dictionary content management has changed a great deal how data are encoded, managed and retrieved, but little work has been done on innovative ways to give end users 'a richer experience'. For more than two decades, computer visualization devices have been set up to present and browse data from a multitude of sources and in many fields, but most current electronic dictionaries (*e-dictionaries*) merely continue to replicate the layout of their traditional printed counterparts to display their contents. Aside from image-based dictionaries that are notorious exceptions (for example: the Merriam-Webster's *Visual Dictionary Online*, QAI's *The Visual Dictionary*), many advantages of computer capabilities for data visualization have yet to be acquired and adapted in this field.

This paper presents the goals, architecture and usability of a prototypical Graphical User Interface (GUI) primarily based on a graph visualization device and used to browse data and discover knowledge through a subset of selected relations that are found in the DiCoInfo (i.e. *Dictionnaire fondamental de l'informatique et de l'Internet*), an online specialized

* We would like to thank M.-C. L'Homme and G. Bernier-Colborne from the OLST for very helpful suggestions and comments on an earlier draft of this paper.

e-dictionary of computing and the Internet. This particular project is part of a larger effort to improve data and knowledge access for language professionals such as technical writers and translators (see L'Homme & Leroyer, 2009; L'Homme et al., 2010).

Its birth is linked to the idea that it was possible to improve the visual and communicative value of dictionary contents using a graph visualization device. First, in displaying the links between the data that appear in field entries: for example, the lexical relationships that exist otherwise among *synonyms*, *derivatives* and *related meanings* of a particular term. Second, in displaying the links between entries that share particular data in some field entry: for example, the relationships among records that mention a particular term as a *derivative* or *related meaning*. Not only do these enhancements seem beneficial, they may be brought together in a single generalized representation that remains neutral with regard to the way the data is accessed. Figure 1 shows the kind of data visualization one can expect to obtain with this approach:

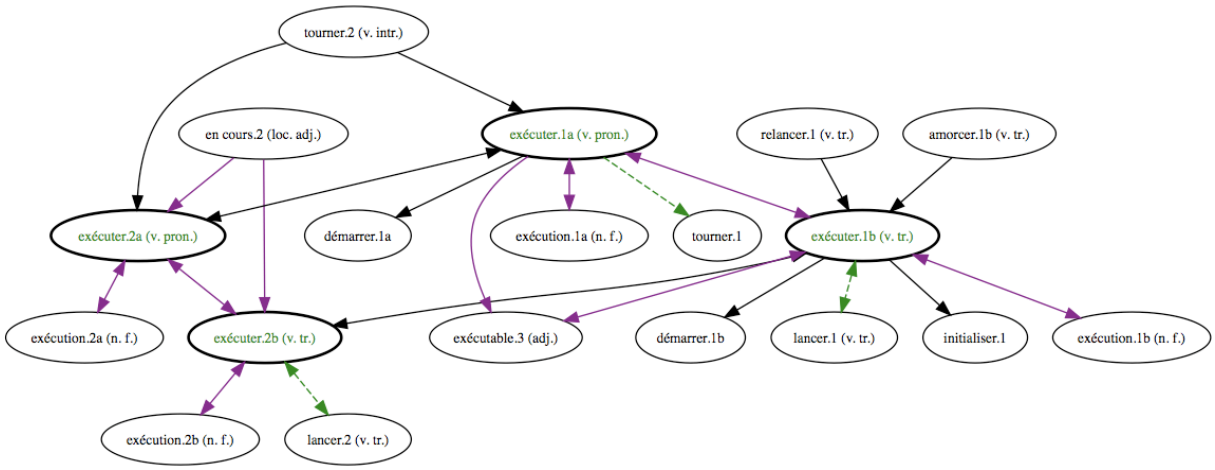


Figure 1: Some of the lexical relations of the polysemous French term 'exécuter'

The actual project was undertaken for two main reasons:

1. We assumed that relationships between terms (perhaps not all, but a large part of them) were likely to be better understood by end users if they were first shown graphically rather than simply listed in tables with textual explanations. In terminology, *taxonomies* and *meronymies* are usually presented in a graphical hierarchy, but other relationships could also lend themselves to a graphical presentation.

2. We also sought to offer a tool for terminologists updating the entries that would help them better assess the consistency of the descriptions. For instance, bidirectional relationships such as *synonyms*, *antonyms*, *derivatives* and *related meanings* could be more easily assessed using a graphical interface.

The rest of the article is organized as follows. Section 2 presents a short overview of traditional GUIs to e-dictionaries and discusses specific drawbacks. It also gives a brief description of a few graph-based GUIs that are found online or downloadable from the Internet. Section 3 first briefly presents the DiCoInfo, and then provides technical details on the architecture and the features implemented so far in our graph-based GUI. Section 4 discusses directions for future work and some of the challenges they raise. Finally, a few concluding remarks are given in Section 5.

2 Old and new ways to explore dictionaries

As previously mentioned, most GUIs to e-dictionaries merely continue to offer traditional outlooks on their contents. Few of them have only the mandatory nomenclature and a display mechanism to view chosen records. Many GUIs offer advanced general and 'by-category' search capabilities that produce (sometimes dynamically) shorter wordlists to help end users access specific contents more efficiently (see for example, *Larousse*, 2011; *Le Petit Robert*, 2011; *OED*, 2011). But wordlists are always presented in the very natural but immutable alphabetically-ordered fashion without showing the links between results. As Manning et al. (2001) mentioned, the basic reason seems to be that, contrary to encyclopedias and thesauri that organize their contents primarily on a conceptual basis, e-dictionaries always compile their contents solely as indexes. Another fundamental reason is simply that they organize and show search results only with respect to field entry organization. A last reason might be that despite the fact that they provide relationships between lexical units, very few encode these relationships formally (however, see Miller, 1993 and Steinlin *et al.*, 2005). As Polguère (2009) puts it, the vast majority are simply *text-based* e-dictionaries, that is they only index field entry data and do not organize them otherwise.

Nonetheless, during the last decade, innovative means for exploring e-dictionaries for end users have been proposed. Some of them rely predominantly on lexical networks and offer appealing and interactive graph visualization devices to navigate within their content (e.g., Jansz et al.'s *Kirrkirr*, 2008; The LexiCon Research Group's *EcoLexicon*, 2009; Thinkmap's *Visual Thesaurus*, 2011; logicalOctopus's *Visuwords*, 2011; Vercruysse's *WordVis*, 2011). However, without appropriate additional control options or display features (like drawing options that allow to select relationship types, see Section 3.2), these GUIs can quickly become confusing and users may have trouble untangling all the information presented.

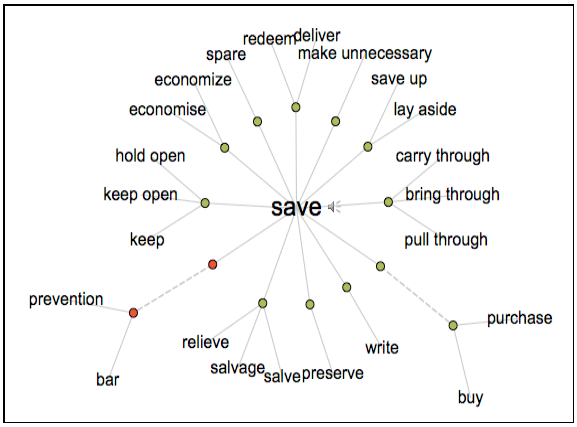


Figure 2
Lexical relations of the English word
'save' from Thinkmap's *Visual Thesaurus*

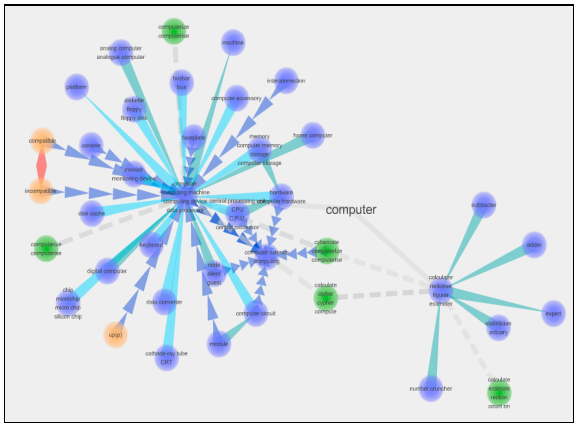


Figure 3
Lexical relations of the English word
'computer' from logicalOctopus's *Visuwords*

3 The DiCoInfo and the DiCoInfo Visuel

As mentioned in the first section, the DiCoInfo is an online e-dictionary that describes terms in the fields of computing and the Internet in French, English and Spanish. It was originally developed as a monolingual tool with the main function of helping end users solve specific knowledge problems associated with this specialized language. From year to year, new

languages and functionalities have been added to assist them with tasks such as translation and text production in a second language (see L'Homme et al., 2009).

3.1 The DiCoInfo terminological database

The records of the DiCoInfo are encoded in XML files that are stored in an eXist database management system (see Meier et al., 2011). Apart from the new graph-based GUI presented in Section 3.2 below, end users access and browse the dictionary contents via two main Web interfaces. The first one, called the *static version*, is a compilation of hyperlinked HTML pages that provides the list of all records in the conventional alphabetical fashion. The second one is a *search version* that mimics a search engine and finds the records containing strings (corresponding to parts of words or terms) in specific field entries such as the usual *headword*, *variants* and *synonyms*, but also in other fields that group different sorts (or *families*) of paradigmatic and syntagmatic lexical relationships. These last relationships are formally classified and encoded by means of the *lexical functions* used in the *Explanatory Combinatorial Lexicology* framework (see Mel'čuk et al., 1995 and Mel'čuk, 1996). Both GUIs are implemented using customary XSLT stylesheets that transform the original XML records and put them together in HTML format (see Clark, 1999).

The next subsection describes the architecture of a new graph-based GUI designed for the DiCoInfo. Technical details are provided on the features that have been implemented so far. It is worth mentioning that subsets of lexical functions used in the DiCoInfo were specifically selected for this first version. These encode paradigmatic relationships, namely *hypernyms*, *synonyms*, *antonyms*, *derivatives* and *related meanings*. *Hyponymic* and *meronymic* relationships are not yet incorporated since the data themselves need to be revised and their drawing polished (see Section 4). Lexical functions encoding syntagmatic relationships are also ignored for now as another strategy for displaying them is presently being developed (see Jousse et al., 2011).

3.2 The DiCoInfo *Visuel*

In its current form, the DiCoInfo *Visuel* is a collection of PHP scripts that carry out the following series of tasks: in addition to generating the welcome and result HTML pages, they manage the search options; query the eXist database; receive and analyze the relational data; and last but not least, generate the graph descriptions (to be sent out to a graph drawing device) with a caption and a hyperlinked index of the terms found in the graph. These tasks may be best sketched as a five-step operational cycle that is summarized in Figure 4.

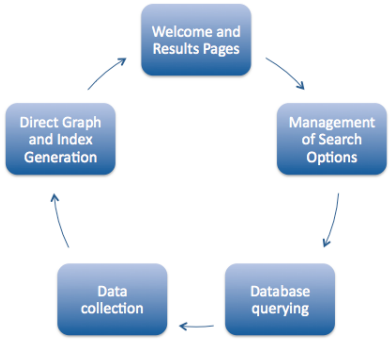


Figure 4: The operational steps in the DiCoInfo *Visuel*

When users access the DiCoInfo *Visuel* without querying it, the main program generates an uncluttered HTML page that welcomes end users (see Figure 5). This page shows usage information, draws the default menu options and finally inserts hyperlinks that point on the one hand to the original HTML static version that contains the word lists of the dictionary; and on the other hand to the sites where the original source code of the Javascript menu framework and the graph drawing tool can be found (respectively, *BlueShoes*, 2011; and

Graphviz, 2011). At this point, users may choose or change some of the search options, type in a query string, and hit the return key for the main program to look up the XML database.

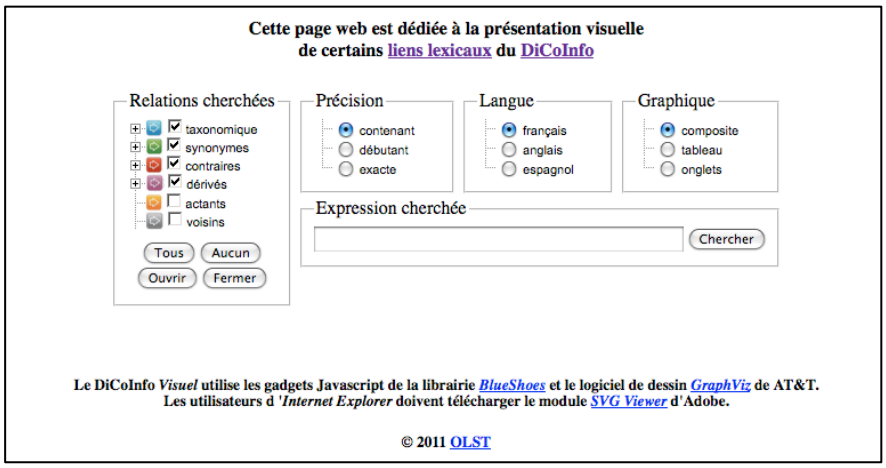


Figure 5: Welcome page of the DiCoInfo *Visuel*

The options menu presents four different groups of options. Shown in Figure 6, the first one deals with the different relationships that may be looked for during searches. Note that these relationship options are grouped in families, as are the lexical relationships in the XML database records. The second menu option allows users to define the search precision. It offers to look for data that matches either partially or exactly the string entered. This option allows users to define their queries according to different needs without having to master regular expressions. The third menu option allows searching different parts of the database depending on the language of the search string. The last option is not implemented yet, but will serve to make different renditions of the results (see Section 4).

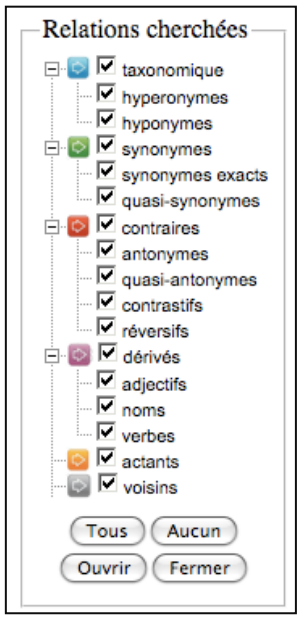


Figure 6: Relationship search options in the DiCoInfo *Visuel*

The next operational step is querying the XML database, but prior to that step the program must put together the queries that are made to the eXist server. One query will be made for each of the (families of) relationships that is selected in the first group of search options. The details of each query are already written in the XQuery language (see Boag et al., 2010) with the exception of the specific values for the search precision, the language option and the searched query string. In other words, for each relationship the program already knows where to look in the dictionary records, and how to format the answer. It is interesting to note that queries for *hypernyms* are recursive and literally walk up and down the relationship paths in the lexical network.

Instantiated XQueries are then submitted to the eXist database server using the XML-RPC protocol (see Scripting News, 2011). For each query, the server returns a collection of very simple XML items of the form: `< link @relation @term1 @term2 />`. Either the searched query string has been found in a *headword*, or in the corresponding relationship field entries. The result items encode the minimal and essential information that the main program needs to know at this point: in the record of '@term1', there exists a relationship of the type

'@relation' that encodes '@term2'. All partial results are then placed in a temporary internal data structure that eliminates the duplicates and records the information to manage the arrowheads of vertices in the future directed graph.

The penultimate operational step is the generation of the graph. The main program first generates its description in the *dot* language (see Ellson, 2011), setting up the display features of the drawing: its general dimensions and orientation; the node list (symbolizing the terms) in which each node has a label, a color, a hyperlink, etc.; and finally the vertices (symbolizing the lexical relationships), each having a color (for its type), a style (for its subtype), its weight and direction, etc. This graph description is then passed to the *Graphviz* drawing software that generates an SVG formatted image (see Dahlström et al., 2011).

Figure 7: Caption for the relations found

The last operational step of the main program is to generate the HTML results page. This page has the same general display as the welcome page, except that the SVG graph is inserted along with a caption for the relations found (see Figure 7) and an index that may be used to access the traditional search GUI of the DiCoInfo.

The following figures exemplify the kind of graph generated by the DiCoInfo *Visuel*: Figure 8 presents a graph obtained with a recursive query that searches for *hypernyms* of the French term 'disque' (Eng. 'disc'); Figure 9 presents *derivatives* found when searching in French for the substring "exéc" (as in 'exécuter', 'exécutable', etc.); Figure 10 shows a part of the graph involving *synonyms*, *derivatives* and *related meanings* among terms containing the substring "program" in English.

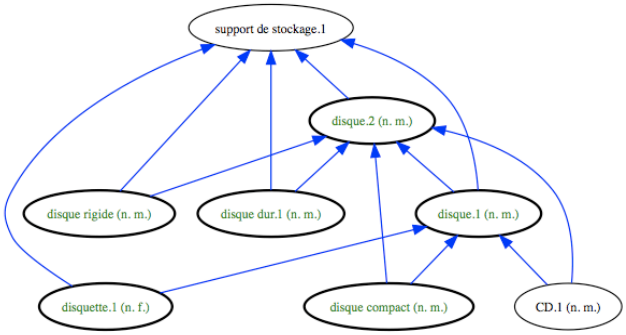


Figure 8: *Hypernyms* of the French terms 'disque'

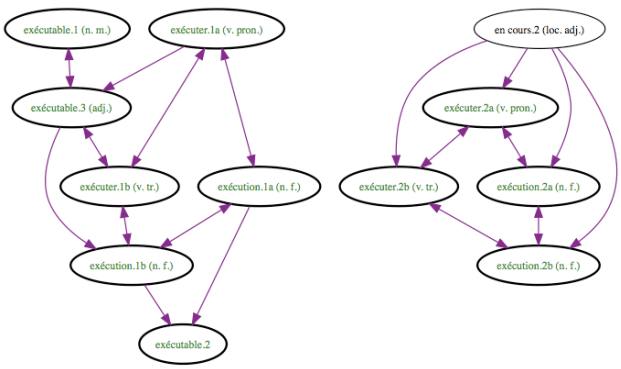


Figure 9: *Derivative* relationships among terms containing the substring "exéc" in French.

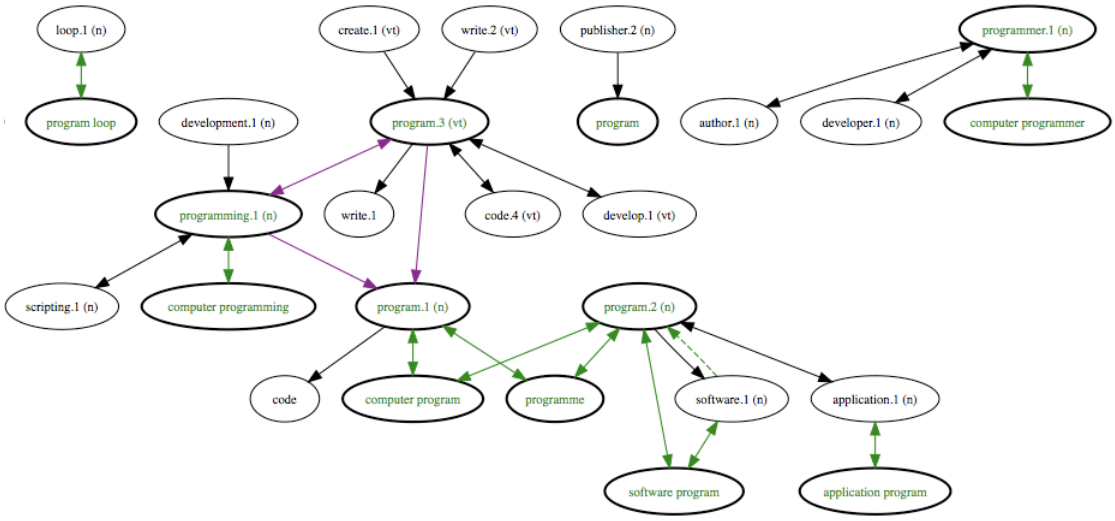


Figure 10: Part of the directed graph with *synonyms*, *derivatives* and *related meanings* of terms containing the substring "program" in English.

4 Improvement and future work

In this section, we discuss three drawbacks noticed during the development and testing of the DiCoInfo *Visuel*. These are dissimilar and will be described independently. For some we propose solutions or improvements. We conclude this section with a brief description of a core feature we intend to implement in the next version.

First, graphs of the DiCoInfo *Visuel* presented so far all have a 'tree' shape, as opposed to the 'spring' shape displayed in other graph-based GUIs mentioned in Section 2. This choice appeared to be a natural one since trees are meaningful and more appropriate for at least subtypes of taxonomic relationships (namely *hypernyms* and *hyponyms*). Other types of relationships seem neutral with respect to this drawing feature. An aesthetical difficulty arises when a particular node has too many direct daughters: these span too widely on the horizontal axis of the tree. One improvement could be made by splitting large sets of daughter nodes into subsets, and distributing them more wisely on the vertical axis with the help of invisible fake nodes. Another solution would be to find the means to mix 'tree' and 'spring' shaped layers in the same graph presentation.

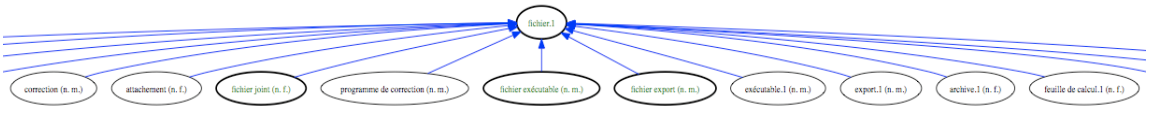


Figure 11: Aesthetical difficulty with nodes having too many direct daughters

Second, as mentioned in Section 2, some queries may simply return too many nodes linked by countless vertices: the entire graph itself becomes extremely difficult to interpret. End users could make a series of more precise queries, but in some cases they may want to visualize all the information anyway. To overcome this 'ergonomic' problem, we plan to implement the last menu options mentioned in Section 3.2 and offer end users the possibility to display the different layers of the resulting graph within table cells or tabs. Another solution would be to display these graph layers in a *Google Gadgets* fashion.

Third, in Section 3.2, we mentioned that setting the search precision option to look for substrings (instead of looking for exact matches) allows to draw richer and more interesting graphs as the XQueries extract large sets of results from the terminological database. Unsurprisingly, this search strategy also finds complex terms by matching their expansion part. For example, a search for 'computer' will locate terms such as 'computer chip', 'computer hacker', 'computer network', and the like. Because shorter terms are preferred as *headwords* and complex ones are mostly encoded as *synonyms*, these terms will appear as orphans if no relationship is found between the *headword* (i.e. 'chip', 'hacker' and 'network') and the base term that corresponds to the expansion (i.e. 'computer').

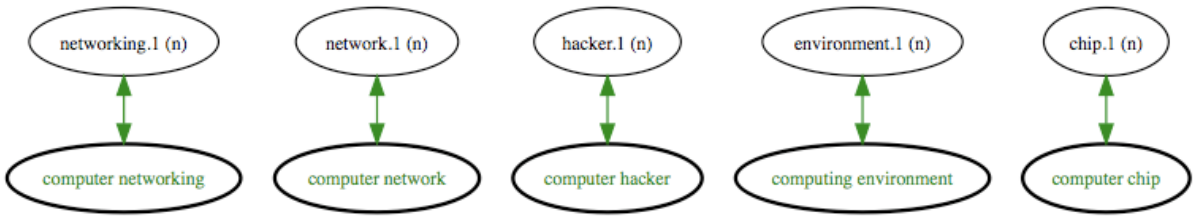


Figure 8: Orphan complex terms found when searching 'computer'

This last problem raises a more significant issue: presently the DiCoInfo *Visuel* is not 'intelligent' and makes no inferences or analogies of any kind. In the next version, in addition to the enhancements discussed above, we intend to build a new architecture of the GUI based on an inference engine as the main program. This new architecture will allow the GUI to draw better graphs, as it will be able to perform the reification of implicit nodes and relationships (see Polguère, 2009). Within this new framework, it will become possible to put in place some inference or analogy mechanisms that will allow generalizing search recursion in the lexical network, and in certain cases compute transitive and deduction closures over the lexical relationships.

5 Conclusion

In this paper, we presented the DiCoInfo *Visuel*, a prototype that puts forward a new method for organizing and visualizing lexical relationships when accessing (specialized) dictionary contents. We addressed the challenge of making dictionary contents accessible and usable for two different types of users (end users and terminologists) through the creation of a single Web GUI. This interface reconciles structured XML data with specialized users' insights when searching, browsing and visualizing terminological information. Our implementation develops a simple and unified solution to the problems of accessing, processing and graphically formatting lexical data in comprehensive ways. Finally, we intend to enhance and expand the software by supplying the actual prototype with an inference engine that we hope will allow to compute lexical analogies and inferences.

Bibliography

Dictionaries

Dictionnaire Le Robert. 2011. *Le Petit Robert de la langue française 2011*. http://pr.bvdep.com/version-1/pr1.asp [last accessed: 25 May 2011]

LexiCon Research Group. 2011. *EcoLexicon, Terminological Knowledge Base*. http://ecolexicon.ugr.es/en/ [last accessed: 25 May 2011]

Éditions Larousse. 2011. *Dictionnaires et encyclopédie en ligne*. http://www.larousse.fr [last accessed: 25 May 2011]

Jansz, K et al. 2008. *Kirrkirr: software for the exploration of indigenous language dictionaries*. http://nlp.stanford.edu/kirrkirr [last accessed: 25 May 2011]

L'Homme, M.-C., 2011. *DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet*. http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi [last accessed: 25 May 2011]

logicalOctopus, 2011. *Visuwords, an online graphical dictionary and thesaurus*. http://www.visuwords.com [last accessed: 25 May 2011]

Merriam-Webster. 2011. *The Merriam-Webster Visual Dictionary Online*. http://visual.merriam-webster.com [last accessed: 25 May 2011]

Oxford University Press, 2011. *The Oxford English Dictionary* (OED). http://www.oed.com [last accessed: 25 May 2011]

QA International, 2011. *The Visual Dictionary*. Éditions Québec Amérique inc. http://www.ikonet.com/en/visualdictionary [last accessed: 25 May 2011]

Thinkmap, Inc. 2011. *The Visual Thesaurus*. http://www.visualthesaurus.com [last accessed: 25 May 2011]

Vercruysse, S. 2011. *WordVis, the Visual Dictionary*. http://wordvis.com/about.html [last accessed: 25 May 2011]

Software

Arn. A. et al. 2011. *BlueShoes: PHP Frameworks & CMS*. http://www.blueshoes.org [last accessed: 25 May 2011]

Boag, S. et al. 2010. *XQuery 1.0: An XML Query Language* (Second Edition) http://www.w3.org/TR/xquery [last accessed: 25 May 2011]

Clark, J. (ed) 1999. *XSL Transformations (XSLT)*. W3C Recommendation. Version 1.0 http://www.w3.org/TR/xslt [last accessed: 25 May 2011]

Dahlström, E. et al. 2011. *Scalable Vector Graphics (SVG)*. Version 1.1 (Second Edition) http://www.w3.org/TR/SVG [last accessed: 25 May 2011]

Ellson, J. et al. 2011. *Graphviz – A Graph Visualization Software*. http://www.graphviz.org [last accessed: 25 May 2011]

Meier, W. et al. 2011. *eXist : an Open Source Native XML Database*. http://exist.sourceforge.net [last accessed: 25 May 2011]

Robichaud, B. 2011. *Le DiCoInfo Visuel*. OLST. Université de Montréal. http://olst.ling.umontreal.ca/dicoinfo/visuel.php [last accessed: 25 May 2011]

Scripting News, Inc. 2011. *The XML-RPC Home Page*.
http://www.xmlrpc.com [last accessed: 25 May 2011]

Other references

Faber, P. et al. 2009. EcoLexicon: A Frame-Based Knowledge Base for the Environment.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. Cambridge MA: MIT Press.

Collins, C. 2008. *WordNet Explorer: Applying Visualization Principles to Lexical Semantics*. Technical Report. Department of Computer Science, University of Toronto, Canada.

Jousse, A.-L., M.-C. L'Homme, P. Leroyer & B. Robichaud. 2011 (to appear). Presenting collocates in a dictionary of computing and the Internet according to user needs. *Proceedings of the 5th International Conference on the Meaning Text Theory*, Barcelona.

L'Homme, M.-C. et al. 2009. *Le manuel du DiCoInfo*. Département de linguistique et de traduction, Université de Montréal.

L'Homme, M.-C. & P. Leroyer. 2009. Combining the semantics of collocations with situation-driven search paths in specialized dictionaries, *Terminology* 15(2), 258-283.

L'Homme, M.-C., P. Leroyer & B. Robichaud. 2010. Advanced Encoding for Multilingual Access in a Terminological Database – A Matter of Balance, In *Terminology and Knowledge Engineering Conference. Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges* (TKE 2010), 12-13 August, Dublin.

Manning, C.D., K. Jansz & N. Indurkhya, 2001. Kirrkirr: Software for browsing and visual exploration of a structured Warlpiri dictionary. *Literary and Linguistic Computing*. 16(1): 123-139.

Mel'čuk, I. 1996. Lexical functions: A tool for the description of lexical relations in the lexicon. In Wanner, L. (ed.), *Lexical functions in lexicography and natural language processing*. Amsterdam/Philadelphia: Benjamins. pp. 37–102.

Mel'čuk, I,, A. Clas & A. Polguère. 1995. Introduction à la lexicologie explicative et combinatoire, Louvain-la-Neuve (Belgique): Duculot / Aupelf - UREF.

Mel'čuk, I. et al. 1984-1999. *Dictionnaire explicatif et combinatoire du français contemporain*. Montréal: Presses de l'Université de Montréal.

Miller, B.P., 1993. What to Draw? When to Draw? An Essay on Parallel Program Visualization. *Journal of Parallel and Distributed Computing*, 18(2): 265-269

Miller, G., R. Beckwith, C. Fellbaum, R. Gross & K. Miller, 1993. Introduction to WordNet: An On-line Lexical Database. Felbaum, C. (ed) *WordNet: An electronic lexical database*. Cambridge MA: MIT Press.

Polguère, A. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*. 43:41-55.

Steinlin, J., S. Kahane, & A. Polguère, 2005. Compiling a ''classical'' explanatory combinatorial lexicographic description into a relational database. *Proceedings of the 2th International Conference on the Meaning Text Theory*, Moscow, pp. 477–485.

Stenmark, D. 1997. *To Search is Great, to Find is Greater: a Study of Visualization Tools for the Web". The result of a class in Human-Computer Interaction*. Internal report.

The unbearable lightness of light verbs. Are they semantically empty verbs?

Sanromán Vilas, Begoña

University of Helsinki
P.B. 24 (Unioninkatu 40 A 516) FIN-00014 Finland
begona.sanroman@helsinki.fi

Track 3: collocations

Abstract

Contrary to the claim that light verbs are semantically empty verbs, this paper defends the viewpoint that the lexical selection of light verbs is based on their meaning (the *hypothesis of the semantic compatibility*). To provide some evidence of this, the paper will concentrate on *pure light verbs* (*tener* 'to have', *hacer* 'to do', etc.) emphasizing the semantic differences between light verb constructions and their verbal counterparts (e.g. *hacer uso* 'to make use' and *usar* 'to use') and pairs of light verb constructions which share the same noun but take a different pure light verb (*dar una excusa* and *poner una excusa* 'to give/make an excuse').

Keywords

Light verbs, light verb constructions, collocations, semantically empty/full verbs, semantic compatibility, Explanatory and Combinatorial Lexicology.

1 Introduction

The aim of this paper is to provide some insights into the nature of light verbs as a first step to elaborate lexicographic entries for these verbs. The term *light verb* (henceforth LV) will be used here to refer to the values of the lexical function (LF) **Oper$_1$**. Within the Explanatory and Combinatorial Lexicology (ECL), a value of **Oper$_1$** is a semantically empty verb (or at least emptied) in the context of its keyword — a predicate — which takes the keyword as its direct object (DO) and the first actant of the keyword as its subject (Mel'čuk, 1996:59). The combination of a LV[1] and its keyword forms a *light verb construction* (LVC), e.g. *to give a talk*, *to have a laugh*. If a LV is a verb which has been emptied of lexical meaning, it follows

[1] We will refer to three types of verbs: LVs or values of **Oper$_1$**, e.g. *dar* 'to give' in *Pedro da un paseo* 'Pedro takes a walk'; the corresponding full verb or heavy verb, *dar* 'to give' in *Pedro da un libro a Ana* 'Pedro gives a book to Ana'; and the verbal counterpart of the LVC, e.g. *pasear* 'to walk' in *Pedro pasea* 'Pedro walks'.

that LVCs such as the preceding ones are equivalent to their verbal counterparts, namely *to talk*, *to laugh*, iff (= if and only if) the noun and the verb have exactly the same meaning.

Contrary to the previous claim regarding LVs as semantically empty verbs, this paper defends the viewpoint that the lexical selection of these verbs is based on semantic grounds. Specifically, we uphold the hypothesis, named the *hypothesis of the semantic compatibility*, that LVs are connected to the noun with which they form a LVC and to the related full verb by means of semantic links, that is to say, a semantic component which is repeated in the LV and the noun, and in the LV and the full verb. E.g. the Spanish noun *elogio* (aprox. 'praise'), which refers to the action and the effect of praising,[2] can combine with two LVs: *hacer* 'to do' (*hacer un elogio* 'to give a praise') and *decir* 'to say' (*decir elogios* 'to say praises'). Far from being synonymous, when *hacer* co-occurs with *elogio*, it relates to the component 'action' of the noun ('X's action expressing approval/admiration about Y's achievements or characteristics). On the contrary, when *decir* combines with *elogio*, it is linked to the component 'effect' ('X's remark expressing approval/admiration about...). Within an aspectual classification, *elogio*, in the sense of 'effect', can be interpreted as an act. Consider the following sentences:

(1) […] diplomático muy distinguido del cual valdría la pena *hacer un elogio* muy extenso…
'a well distinguished diplomat of whom it would be worth to praise extensively…'

(2) ¿Un hombre que *dice muchos elogios* cuando apenas te conoce, puede ser un buen mentiroso?
'A man who pays you many compliments when he has just recently met you, is he a good liar?'

On the one hand, *elogio* 'praise' in (1) refers clearly to an action 'what a person voluntarily does during a period of time' because it has duration. The intensifier adjective *extenso* 'X is extensive' means that 'X occupies time, X lasts'. On the other hand, *elogio* in (2) refers to an act 'what a person voluntarily and punctually does' because it has no duration; instead of that, it can be quantified as is seen in (2), where *elogio* is in the plural form. The semantic links these two LVs — *hacer* in *hacer un elogio* and *decir* in *decir elogios* — keep with the corresponding heavy ones are: 1) LV *hacer* shares the component 'action' with the corresponding heavy verb *HACER*, a polysemous verb which has several LUs linked by the same 'action' component; 2) LV *decir* shares the complete meaning of *decir* as a full verb 'to express by words', that we will simplify here as 'speech act'.

This paper will concentrate in *pure LVs* (Alonso Ramos, 2004:91), a subclass within LVs which has a greater degree on meaning extension than the others. Pure LVs are considered to have completely lost their lexical meaning till such an extent as to become verbs with only grammatical meaning like auxiliary verbs. This paper will provide some evidence to show that even pure LVs have their own lexical meaning. Firstly, LVCs, e.g. *hacer uso* 'to make use', will be contrasted to their verbal counterparts, e.g. *usar* 'to use' — a verb with the same meaning as the noun and morphologically derived from it — to emphasize the differences. Secondly, pairs of LVCs sharing the same noun but taking a different pure LV (*dar una cabezada*, *echar una cabezada* 'to have/take a snooze', both equivalent to *cabecear* 'to have/take a snooze') will be compared to highlight also the dissimilarities. In both cases the

[2] *Elogio* can be defined as a lexical unit (LU) with a disjunctive semantic component (*elogio*: 'action or effect') or as a vocable with two LUs each of them with one of the components (*elogio**1***: 'action' and *elogio**2***: 'effect').

relevant question is: what is the semantic contribution of the pure LV to the LVC when the meaning of the LVC does not match up exactly with its verbal counterpart or when the meaning of two apparently synonymous LVCs do not coincide?

This research is mainly carried out within the theoretical and methodological framework of the ECL (Mel'čuk et al., 1995), which is part of the Meaning Text Theory (MTT) (Mel'čuk, 1997). Data has been collected from two corpora — *Corpus de referencia del español actual* (*CREA*) and *Corpus del español* (*CdE*) —, several Spanish monolingual dictionaries and different Internet pages. The paper is organized in five sections. After this Introduction, a review of some approaches to lexical meaning in LVs will be offered (Section 2). Next we will prove the presence of meaning in LVs by comparing two types of expressions: LVCs and their verbal counterparts (Section 3.1), and pairs of LVCs sharing the same noun equivalent to a verbal counterpart (Section 3.2). In Section 4 some tentative generalizations will be made concerning the meaning of LVs. Finally Section 5 will draw some conclusions.

2 The nature of the meaning of LVs

In the ECL, a LVC is considered a type of *collocation* (Mel'čuk et al., 1995:46), where the collocative, or LV, is lexically selected by its base, the noun, in a more or less arbitrary way to acquire a sentence configuration. As there is no lexical meaning in the LV, its role is restricted to add grammatical information about tense, mood and person (Alonso Ramos, 2004:24). Despite general claims denying that LVs are meaningful LUs, some concessions, within ECL, have been done. Accordingly, Mel'čuk (1992:32-33) has remarked that even if values of **Oper$_1$** are semantically (quasi-)empty, LUs included as values of the LF for the same keyword are not necessarily exact synonyms. Furthermore, the author declares that they can differ semantically from each other in many nuances. In the same line, Alonso Ramos (2004:85-87) has drawn attention to two possible perspectives to analyze the semantic emptiness of LVs: from the syntagmatic viewpoint, she claims that LVs are semantically empty$_2$ because they are not selected for their lexical meaning, and consequently, they can only provide the noun with temporal, modal, etc. information. From the paradigmatic viewpoint, LVs can be semantically empty$_1$ in the sense that they have an abstract meaning including only generic components which characterize the semantic class of the verb. Further on, Alonso Ramos (2004:91-93) has recognized that not all LVs are semantically empty$_1$, some of them maintain semantic links with other senses of the corresponding full verbs.

Apart from the previous authors, who admit that LVs can be somewhat meaningful even if this is not the focus of their approach, it is possible to find many other researchers who have addressed the question of the lexical meaning of LVs in a more direct way. Among these authors, we can distinguish two different approaches: 1) those who concentrate on the relationship between LVs and their heavy counterparts, and 2) those who lay emphasis on the relationship between LVs and the keyword noun.

1) The focus on the relationship between LVs — *to take* in *X takes a walk* — and their heavy counterparts — *to take* in *X takes an apple from the tree* — brings together mainly authors adhering to cognitive frameworks. In general lines, the semantic links between both LUs are considered as extensions of a polysemous word. In cases where they cannot manage to isolate a distinct meaning of a particular LV, it is assumed that the meaning of the LV has to be examined within the construction where it is placed. As representative works in this approach,

we mention the analysis of *to take* (Norvig & Lakoff, 1987) or the extensive study about *to give* (Newman, 1997). Brugman (2001), for her part, has reviewed several LVs in order to verify that they have meaning and not only function, as well as to prove that the relationship between light and heavy verbs shows regularities cross-lexically, that is to say, the same semantic links which relate a LV to the heavy one is repeated in other pairs.

2) On the contrary, it seems that authors whose target is to find an explanation for the relationship between the LV and the noun occupying the position of DO belong mostly to functional approaches, particularly associated to lexicographic projects. Among them, we refer to Apresjan (2007, 2009), Barrios (2010), Bosque (2004) or some authors within the framework of the "lexique-grammaire" such as Giry-Schneider (1987) and Vivès (1984). They agree that aspectuality plays a major role in the relation LV-noun. According to Apresjan (2009), the selection of a LV is conditioned by the lexical meaning of the LV and the semantic class of a Vendlerian classification to which the noun belongs. He claims that there is "semantic agreement" between the LV and the noun: "they should have at least one non-trivial recurrent (repetitive, common) semantic component in their meaning" (Apresjan, 2009:4). Bosque talks about "redundancy" and "agreement of lexical features" (Bosque, 2004a:47, 2004b).

Beside these approaches,[3] Reuther (1996:198-199) proposes the description of the meaning of a LV in three different parts: a more general (taxonomic) part, a specific part which shows the semantic links with other senses of the given verb, and another specific part containing the semantic characteristics of nouns that typically appear in collocations with the verb.

3 An analysis of pure LVs within LVCs

In this Section, we will compare LVCs with their verbal counterparts, on the one hand, and pairs of LVCs, on the other hand. In both cases, there is a monolexical verb semantically related to the LVC; LVs belong to the group of pure LVs. We will try to prove that LVs contribute meaning to LVCs and to some extent, we will show the semantic links between ligth-full verbs and LVs-nouns. Pure LVs are a special group of LVs frequently compared to (quasi-)auxiliaries because of their semantic emptiness. In Colloquial Spanish studies, they have been called *verba omnibus* or *pro-verbs*; in language teaching, they are often used to improve the vocabulary asking students to replace LVs with more precise equivalents, e.g. *verter* instead of *dar la opinión* 'to give the opinion'. Although there is no complete agreement about the list of pure LVs in Spanish, scholars accept that, at least, *hacer* 'to do/make', *dar* 'to give', *tener* 'to have', *poner* 'to put', *tomar* 'to take' and *echar* 'to throw' belong to the group.

3.1 LVCs *versus* verbal counterparts

The equivalence between LVCs, such as *tener respeto* 'to have respect' or *dar una orden* 'to give an order', and verbal counterparts, *respetar* 'to respect' or *ordenar* 'to order', can be based on the fact that pure LVs are semantically empty verbs. In this sense, the semantic weight of

[3] This classification is not exhaustive. Some other researchers have addressed the question in more or less depth, as for example Martín Mingorance (1998) or de Miguel (2007) for the Spanish language.

the construction is carried by the predicative noun. It cannot be denied that in many contexts, the verbal counterpart and the LVC itself are equivalent expressions. E.g. *ordenar* 'to order' and *dar una orden* 'to give an order' can replace each other in (3):

(3) a. […] el teniente Mauricio *dio la orden* de atacarlos.
'[…] the lieutenant Mauricio gave the order to attack them.'
b. […] el teniente Mauricio *ordenó* atacarlos.
'[…] the lieutenant Mauricio ordered to attack them.'

Nevertheless, the replacement is not possible in all the situations. For instance, (4a) can be expressed with *dar una orden* 'to give an order' but not with *ordenar* 'to order' (4b):

(4) *a*. […] cuando un tío *da una orden*, se cumple a rajatabla.
'[...] when an uncle gives an order, it must be carried out to the letter.'
b. […] cuando un tío *ordena* *(algo), se cumple a rajatabla.
'[...] when an uncle orders *(something), it must be carried out to the letter.'

The reason is that LVCs accept not to state explicitly the content of the object — e.g. *orden* 'order' in (4a) —, putting the emphasis in its mere existence. However, with verbal counterparts, it is needed to specify the content unless a pronominal mark has been left instead, as *algo* 'something' in (4b). On the contrary, it seems that verbal counterparts (5a), and not LVCs (5b), are the only ones with the capacity to form performative statements. For instance, in (5a), the speaker, talking in the first person singular of the Indicative Present, is performing a speech act in the moment of the enunciation, while the same speaker in (5b) only describes a speech act.

(5) *a*. […] les *ordeno* que guarden silencio.
'[...] I order you to keep silent.'
b. ?Les *doy la orden* de que guarden silencio.
'[...]?I give you the order to keep silent.'

Differences between pure LVCs with their verbal counterparts can be classified in three groups (Sanromán Vilas, 2009): 1) those which depend on the item we are dealing with, either the simple verb or the verb + noun phrase; 2) those which are based on the semantic class of each of the LUs: the LV, the predicative noun or the correlated single verb; and 3) those which rely on the particularities of each LU, e.g. *dar* 'to give' as a concrete LV, *broma* 'joke' as a predicative noun and *sospechar* 'to suspect' as a correlated single verb. The first type of differences should be addressed in the Grammar because of their general and systematic nature. However, the second and third types should be registered in the Dictionary because of their lexical-semantic character. For reasons of space we only offer an overview of the differences of the first type.

The fact that we are dealing with two different grammatical units — a verb and a phrase — gives rise to four major differences between these expressions. Firstly, verbal counterparts of LVCs happen to be often polysemous words and LVCs refer often to one of the senses of the verb. E.g. *usar* 'to use' can display meanings like: 'to make useful for a particular purpose' in (6a); 'to wear on usually' in (6b); 'to get maximum benefit from something in a particular moment' in (6c).

(6) *a*. [...] es un producto que *se usa* en la agricultura andaluza...
'[...] it is a product used in the andalusian agriculture'
b. El abrigo, porque allá no *se usa* sombrero ni nada de eso.

'The coat, because there neither hat nor nothing like that is used'
c. Nuestro gobierno haría bien en [...] *usar* todos los medios a su alcance...
'Our government would do well by [...] using all the means within its reach'

Among these meanings, the only one which can be expressed by *hacer uso* 'to make use' is (6c): *Nuestro gobierno haría bien en [...]* hacer uso *de todos los medios*... 'Our government would do well by [...] making use of all the means...'. Secondly, using the verb, the actual moment of the action can be specified as an act — or as a series of reiterative acts — (*El conferenciante* tose *discretamente* 'The lecturer coughs discreetly'); however, with the LCV, it is a characteristic of the action, or act, as a whole what is described (*El conferenciante* tiene una tos *discreta* 'The lecturer has a discreet cough'). One way of characterizing an act is by counting the number of times it is repeated (*Le dio dos/tres besos* 'He gave her two/three kisses') or by specifying the type (*Le dio un beso fraternal/de saludo* 'He gave a brotherly/greeting kiss'). Thirdly, predicative nouns and verbal counterparts of LVCs (7) often have the same number of Sem(antic) actants (X, Y, Z), which can be expressed as S(urface)-Synt(actic) actants depending either on the noun or on the verb.

(7) *a*. *la tos de X* 'X's cough' — *X tose* 'X coughs'
b. *el uso de X de Y* 'X's use of Y' — *X usa Y* 'X uses Y'
c. *la orden de X a Y de Z* 'X's order to Y to Z' — *X ordena a Y Z* 'X orders Y Z/Z to Y'

However, when predicative nouns make part of LVCs, most of their Sem actants and the noun itself become dependents on the LVs in SSyntR(epresentation). Thus, in LVCs in (8), the first Sem actant of the noun (X) is always the subject of the LV; when the noun has two Sem actants, Y can be expressed in SSyntR as dependent either on the noun, as *influencias* in (8b), or on the LV, as *su padre* in (8c); when the noun has three Sem actants, as in (8d), Y (*los estudiantes*) is expressed as dependent on the LV and Z (*de que regresen*...), on the noun.

(8) *a*. El barbero *tiene tos*. 'The barber has cough'
b. Carlos *hizo uso* de sus influencias. 'Carlos made use of his contacts'
c. Carlos le *tiene respeto* a su padre. 'Carlos has respect towards his father'
d. El Gobierno les *da la orden* a los estudiantes de que regresen a sus clases. 'The Government gives the order to the students to go back to their lessons'

Lastly, a LVC and the verbal counterpart do not have the same Sem-Comm(unicative) S(tructure).[4] In MTT (Mel'čuk, 2001) a Sem-CommS describes how the utterance's propositional meaning is organized to be transmitted as a message. For instance, the propositional meaning represented in (9) can be expressed by a single verb (10b) or by a LVC (10a):

(9) 'X applies soap and jets of water all over X's body to wash X's body'

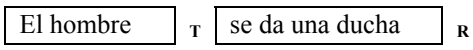
(10) *a*. El hombre *se da una ducha*. 'The man takes/has a shower'
b. El hombre *se ducha*. 'The man showers himself'

To organize the propositional meaning of an utterance several Sem-Comm oppositions have to be considered. For the purpose of this explanation, only two are needed: thematicity and unitariness. Thematicity, the main organizational feature of any message (Mel'čuk, 2001), is about the basic division of the propositional meaning in Theme and Rheme. We say "basic"

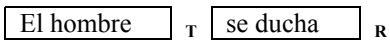
[4] A similar remark is made within the Functional Lexematic framework (Martín Mingorance 1998:22).

because in any message there is always something to say (Rheme) about something (Theme). Sentences in (10) are divided in Theme ($_{T}$) and Rheme ($_{R}$) in (11). The difference between (11a) and (11b) is that the former, but not the latter, admits a second-order division in Theme and Rheme, presented in an approximately way in (12):

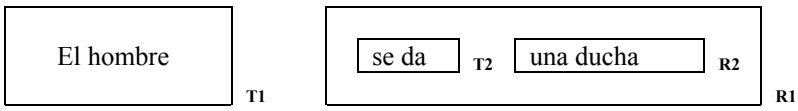
(11) *a.*



b.



(12)



According to Mel'čuk (2001, forthcoming), thematicity has consequences in the lexicalization of meaning in the sense that Sem nodes that are expressed together by one LU should not belong to different thematic areas. As can be seen, the meaning in (9) has lexicalized in (11b) in one LU (*ducharse*) because its Sem nodes belongs to the same thematic area ($_{R}$); on the contrary, (9) is expressed in two LUs (*dar* and *ducha*) in (12) because the Sem nodes pertain to different Sem areas ($_{T2}$ and $_{R2}$). With regard to unitariness, Mel'čuk (2001) refers to the important distinction made in natural languages between representing a situation as a single fact (called *unitary*) or as a simultaneous or subsequent occurrence of several facts (named *articulated*). The simple verb *ducharse* describes a single fact, while the LVC *dar una ducha*, among other nuances, is used to talk about more than one fact which take place simultaneously or consecutively.

3.2 Pairs of LVCs equivalent to a single verbal counterpart

There are LVCs which share the same predicative noun, although they can admit two (or even three) pure LVs: *hacer/tomar un descanso* 'to get/have/take a rest', *dar/poner una excusa* 'to give/make an excuse', etc. Within a linguistic framework assuming the semantic emptiness of pure LVs, the equivalence of these pairs of expressions is clear. And it becomes even clearer if, in addition to this, it is shown that for each pair of expressions, a correlated single verb is available (*descansar* 'to rest', *excusarse* 'to excuse', etc.). Actually, it is possible to find language samples where a pair of LVCs sharing the same noun are synonymous (13):

(13) Voy a *echar/dar una cabezada*. ¿Quieres tú dormir también un poco?
'I am going to take/have a snooze. Do you want also to sleep a little bit?'

However, (14) proves that these expressions are not equivalent in all the contexts. In (14) *dar una cabezada* cannot be replaced by *echar una cabezada*:

(14) Sin darse cuenta, *dio/*echó una cabezada* de la cual se levantó un poco sobresaltado...
'He took a snooze without noticing it and he woke up startled'

The reason is that while *cabezada* 'a short and light sleep' has a neutral semantic component regarding the volitional character of the sleep, what means that the noun can refer either to voluntary or unvoluntary sleeps, *echar* co-occurs only with voluntary acts, and *dar* can select either voluntary or unvoluntary acts. The opposition *dar/echar una cabezada*, far from being an isolated case, represents an example of the links of semantic compatibility between LVs and nouns. In what follows, some examples within the semantic field of emotion nouns,

speech nouns and nouns of corporal acts will be shown. Within the semantic field of emotion nouns two groups can be distinguished (Sanromán Vilas, 2003): internal cause emotion nouns (ICs) and external cause emotion nouns (ECs). In ICs, as *admiración* 'admiration', *envidia* 'envy', *odio* 'hate', the emotion is born in the experiencer as a result of a judgment made about an object. With ECs, as *alegría* 'joy', *asombro* 'amazement', *disgusto* 'upset', it is required the existence of an external fact that triggers the emotion. Overall, ICs co-occur with *tener* 'to have' and *sentir* 'to feel'[5]. When ICs combine with ***tener1*** (15), it is emphasized that these nouns behave mainly as emotional attitudes addressed to an object; that is why the expression of the object is compulsory (*Tolstoi* in (15a), *le* in (15b)).

(15) *a*. *La admiración que tuvo* Pasternak por Tolstoi —a quien conoció de niño...
'The admiration Pasternak had towards Tolstoi — to whom he met as child...'
b. Especialmente Primitivo le *tenía envidia* porque amaba el poder...
'Specially Primitivo had envy towards him because he loved power...'

However, if ICs come with *sentir*, the stress is on the feeling. So, the role of the experiencer is foregrounded (16) while the object of the emotion is presented as background information:

(16) Comencé a *sentir respeto y cariño* por aquel estudiante...
'I started to feel respect and affection to that student'

ECs can co-occur also with *tener* and *sentir*. *Sentir* points at the experience too, but *tener* emphasizes the nature of stage-level predicates of ECs and exhibits a different government pattern with ECs than it does with ICs (Sanromán Vilas, 2008). Contrast (17), where ***tener2*** has two SSynt actants (the experiencer and the emotion), with (15), where ***tener1*** governs three SSynt actants: the experiencer — *Pasternak* (15a), *Primitivo* (15b) —, the emotion — *admiración* (15a), *envidia* (15b) — and the object of the emotion — *Tolstoi* (15a), *le* (15b) —.

(17) Cuando murió la madre, *tuvo una depresión* tremenda.
'When the mother died, he had a deep depression'

Moreover, if ECs are so momentary that the cause of the emotion and the emotion itself become identified the one with the other, ECs can co-occur with *llevarse* lit. 'to take away'. Notice that in (18), *llevarse* can combine with *disgusto* 'upset', *susto* 'fright', *alegría* 'joy', but not with *desesperación* 'desperation' or *angustia* 'anguish':

(18) [...] me *llevé un disgusto* (*susto*, *alegría*, **desesperación*, **angustia*) cuando mi hija mayor decidió casarse a los 19...
'I got upset (I had a fright, joy, desperation, anguish) when my oldest daughter decided to marry at 19...'

Speech nouns such as *énfasis* 'emphasis', *excusa* 'excuse', *objeción* 'objection', etc. can combine with more than one pure LV. We will present briefly ***excusa2*** 'excuse' = 'pretext that X addresses Y to do or not to do Y' and *objeción* 'objection' 'argument that X expresses against Y'. ***Excusa2*** 'excuse' co-occurs with *poner* 'to put' and *dar* 'to give': when the selected LV is *poner* 'to put', the accent lies on the content of the excuse; however, when it is used with *dar* 'to give', the communicative side of the excuse, addressed to somebody, is emphasized. This is the reason why *excusa*, with *dar*, accepts the adjective *pública* 'public' (19a), but the use of the same modifier with *poner una excusa* 'to put an excuse' is of doubtful acceptability (19b).

[5] *Sentir* 'to feel' is not considered an empty **Oper$_1$** but with emotional/perceptual meaning.

(19) *a*. [...] quienes no han tenido la gentileza [...] de *dar una excusa* PÚBLICA para justificar...
'[...] who has not been kin enough [...] to give a public excuse to justify...'
b. [...] quienes no han tenido la gentileza [...] de *poner una excusa* *PÚBLICA...
'[...] who has not been kin enough [...] to make a public excuse...'

Objeción 'objection', which accepts *hacer* 'to make' and *poner* 'to put', comes preferably with *hacer* when the focus is on the action (20a) and with *poner*, when the target is the result (20b).

(20) *a*. [...] *la objeción se hizo* de forma atenta y respetuosa...
'[...] the objection was made kindly and respectfully...'
b. Solamente le tengo que *poner una objeción* a este rico postre...
'I only must make an objection to this nice dessert...'

Concerning acts and actions related with the body (*cabezada* 'snooze', *descanso* 'rest', *ducha* 'shower', etc.), we will comment *descanso* 'X's pause during an activity' and *ducha* 'shower' 'X's act, X applying soap and jets of water all over X's body to wash X's body'. Both nouns co-occur with *tomar(se)* 'to take' which focus on X, Agent and Patient of the action (21):

(21) *a*. [...] para que los empleados *tomaran sus descansos* y comieran...
'[...] for the employees to take their rests and eat...'
b. *Me tomé una ducha* caliente para relajarme...
'I took a hot shower to relax myself...'

Descanso 'rest' combines also with *hacer* 'to make' and *ducha* 'shower' with *darse* 'to give to oneself'. *Hacer un descanso* 'to have a rest' focuses on the rest itself as the period when X is not doing the main activity (22). *Darse una ducha* 'to have a shower' gives prominence to X, not as a Patient or Beneficiary of the action, but as an Agent of the action itself (23).

(22) [...] un agradable jardincillo con sombra, que constituye un buen lugar para *hacer un descanso*.
'[...] a small and nice garden with shadow, which represents a good place to have a rest'

(23) Se levantó, *se dio una ducha* y bajó a cancelar la cuenta.
'He woke up, had a shower and went down to cancel the bill'

4 Tentative generalizations about LVs

Pure LVs share lexical aspect and some characteristics of the subject, e.g. volitionality, with predicative nouns. In some cases these are also the semantic features LVs have in common with the heavy counterparts, e.g. *hacer* 'to do/make'. In other cases, possible nuances in the lexical aspect of LVs are supported by its government pattern, e.g. *tener* 'to have'. Nevertheless, LVs like *tomar* 'to take', *sacar* 'to take out' o *poner* 'to put' inherit the basic meaning of the heavy counterpart by means of metaphorical links. For instance, the basic meaning of *tomar* 'to take' is 'to grasp with the hand'. The relation between *tomar* 'to take' as a heavy verb and as a LV is not so evident in expressions such as *tomar una ducha* 'to take a shower' or *un trago* 'a sip', unless it is claimed that in both situations hands are used. The semantic link is still less visible in *tomar la siesta* 'a nap' or *una decisión* 'a decision'. Thus, we consider that what LV *tomar* inherits from its heavy counterpart 'to grasp with the hand' is the fact that the direction of movement is always towards the self. This kind of deictic information, withheld by the LV, makes possible that *tomar* is mainly used for reflexive actions, that is, actions where Agent and Patient are the same person.

In aspectual classifications, there is a basic division between events with duration and without it. Here we have found a correlation between this division and the use of LVs. Nouns which combine with *tener* and *llevar*(*se*) are mainly states and never admit *dar*, *hacer* or *echar*, which come with actions, acts or activities. Regarding *tener* and *llevarse*, two generalizations can be done. First, *tener* is a polysemous LV that has at least two LUs: *tener***1**, with three Synt actants, combines with individual states, mainly attitudes; *tener***2**, with two, co-occurs with episodic states. Second, *llevarse* combines only with episodical states, mainly those which have a very short duration. *Hacer*, *dar*, *echar*, *tomar*, *poner* co-occur with actions and acts. As said above, *tomar* goes with reflexive actions, while *poner* prefers results of actions. *Hacer* and *dar* are neutral with respect to the volitionary dimension; however, *echar* is marked as [+ volitionary]. Finally, *hacer* can combine with activities, but not *dar* or *echar*. We consider that this type of generalizations, once tested enough, should be consigned in the Dictionary under the entry of the corresponding light verb.

5 Conclusions

In this paper we have tried to show that LVs are not semantically empty, but they have links of semantic compatibility with the noun within the same LVC and with the heavy counterpart. To prove this hypothesis, we have shown that there is no complete equivalence between LVCs (*dar una orden*) and their verbal counterparts (*ordenar*), and between two LVCs sharing the same predicative noun (*hacer/poner una objeción*). Within the limitations of this study, some preliminary and tentative generalizations about LVs have been done. The study will be developed further with the aim of elaborating lexicographic entries for Spanish LVs.

Acknowledgements

I would like to thank the three reviewers of this paper whose comments, objections and corrections have helped to improve the text.

Bibliography

Alonso Ramos, M. 2004. *Las construcciones con verbo de apoyo*. Madrid: Visor Libros.

Alonso Ramos, M. 2004. *Diccionario de colocaciones del español* http://www.dicesp.com.

Apresjan, J. & M. Glovinskaja. 2007. Two Projects: English ECD and Russian Production Dictionary. In *Proceedings of the 3rd International Conference on MTT*. München: Wiener Slawistischer Almanach <http://meaningtext.net/mtt2007/proceedings> (22/01/2011).

Apresjan, J. 2009. The Theory of Lexical Functions: An Update. In *Proceedings of the Fourth International Conference on MTT*. 1-14. Montreal: OLST. http://olst.ling.umontreal.ca/pdf/ProceedingsMTT09.pdf (28/05/2010)

Barrios Rodríguez, A. 2010. *El dominio de las funciones léxicas en el marco de la teoría sentido-texto*. http://ddd.uab.cat/pub/elies/elies_a2010v30/index30.html (02/06/2010).

Bosque, I. 2004a. La direccionalidad en los diccionarios combinatorios y el problema de la selección léxica. In Cabré, T. (ed). *Linguistica teórica*: *análisi i perspectives*. 13-58. Barcelona: Universitat Autònoma Barcelona.

Bosque, I. 2004b. *REDES*. *Diccionario combinatorio del español contemporáneo*. Madrid: SM.

Brugman, C. 2001. Light Verbs and Polysemy. *Language Sciences* 23:551-578.

CdE: Davies, M. *Corpus del español* http://www.corpusdelespanol.org/x.as.

CREA: RAE: *Corpus de referencia del español actual* http://www.rae.es.

Giry-Schneider, J. 1987. *Les prédicats nominaux en français. Les phrases simples à verbe support*. Genève: Droz.

Martín Mingorance, L. 1998. Las unidades sintagmáticas verbales en inglés y en español. Metodología de análisis. In Marín Rubiales, A. (ed). *El modelo lexemático-funcional. El legado lingüístico de Leocadio Martín Mingorance*. Granada, Universidad de Granada, 19-31.

Mel'čuk, I. 1992. Paraphrase et lexique. La théorie Sens-Texte et le Dictionnaire explicatif et combinatoire. In Mel'čuk, I. et al., *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexicosémantiques III*. 9-58. Montreal: Université de Montréal.

Mel'čuk, I. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Wanner, L. (ed). 37-102.

Mel'čuk, I. 1997. *Vers une linguistique Sens-Texte*. Paris: Collègue de France.

Mel'čuk, I. 2001 *Communicative Organization in Natural Language*. Amsterdam: Benjamins.

Mel'čuk, I. (forthcoming). *Semantics: From Meaning to Text*. Amsterdam: Benjamins.

Mel'čuk, I., A. Clas & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve, Duculot.

Miguel, E. de 2007. El peso relativo de los nombres y los verbos: cambios, ampliaciones, reducciones y pérdidas del significado verbal. In Delgado Cobos, I. & A. Puigvert Ocal (eds). *Homenaje a Ramón Santiago I*. Madrid : Ediciones del Orto, 295-326.

Newman, J. 1996. *Give: A Cognitive Linguistic Study*. Berlin: Mouton de Gruyter.

Norvig, P. & G. Lakoff. 1987. Taking: A Study in Lexical Network Theory. *Proceeding of the 13th Annual Meeting*. 195-206. Berkeley: Berkeley Linguistics Society.

Reuther, T. 1996. On Dictionary Entries for Support Verbs: The Cases of Russian *vesti*, *provodit'* and *proizvodit*". In Wanner, L. (ed). 181-208.

Sanromán Vilas, B. 2003. *Semántica, sintaxis y combinatoria léxica de los nombres de emoción en español*. Helsinki: Yliopistopaino.

Sanromán Vilas, B. 2008. El verbo *tener* como marcador aspectual de los nombres de emoción. *Español Actual* 89:99-110.

Sanromán Vilas, B. 2009. Diferencias semánticas entre construcciones con verbo de apoyo y sus correlatos verbales simples. *ELUA* 23:289-314.

Vivès, R. 1984. L'aspect dans les constructions nominales prédicatives: *avoir*, *prendre*, verbe support et extension aspectuelle. *Linguisticae Investigationes* 8:161-185.

Wanner, L. (ed). 1996. *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam: Benjamins.

Barcelona, 8-9 September 2011

Towards a lexicographic description of the Russian verb *терпеть*

Alexei Shmelev

(1) Moscow Pedagogical State University
M. Pirogovskaia 1, Moscow
(2) Institute of Russian Language, Russian Academy of Sciences
Volkhonka 18/2, Moscow
shmelev.alexei@gmail.com

Abstract

The paper provides a lexicographic portrait (or, rather, a rough sketch for a lexicographic portrait) of the Russian verb *терпеть* along the lines suggested in (Шмелев, 2003). Special attention is paid to the aspectual behavior of the verb and to its numerous derivatives since it is morphologically very productive and different derivatives correspond to its different lexical meanings (in addition, some of the connotations of the verb may become a part of the lexical meaning of a derivative). A brief overview of the lexical meanings of the verb follows:

1.1 ‘to bear suffering with patience’
1.2 ‘to resist impatience’
2 ‘to tolerate’
3 OPER (‘an occurrence of something bad’)

In addition, the cultural significance of the attitudes encoded in various readings of the verb *терпеть* and its derivatives is discussed.

Keywords

Lexicography, polysemy, cultural semantics

1 Introductory remarks

In the paper, I will provide a lexicographic portrait (or, rather, a rough sketch for a lexicographic portrait) of the Russian verb *терпеть* along the lines suggested in (Шмелев, 2003). I will pay special attention to the aspectual behavior of the verb and to its numerous derivatives since it is morphologically very productive and different derivatives correspond to its different lexical meanings (in addition, some of the connotations of the verb may become a

part of the lexical meaning of a derivative). A brief overview of the lexical meanings of the verb follows:

1.1 'to bear suffering with patience'
1.2 'to resist impatience'
2 'to tolerate'
3 OPER ('an occurrence of something bad')

In addition, I will discuss the cultural significance of the attitudes encoded in various readings of the verb *терпеть* and its derivatives. It is sometimes claimed that the attitude described as *терпеть* is very salient in Russian culture because a long history of oppressive rule in combination with a severe climate have led to the idea that the ability to endure suffering is the only way to survive (Gladkova, 2007, p. 143-144). Whether or not we accept this claim, one should agree that *терпеть* is a very common language-specific verb and deserves a detailed semantic description.

2 *Терпеть 1* (sufferance, patience)

2.1 Терпеть 1.1

2.1.1 Explication

X терпит (Y) 'A person X experiences something bad (Y) against X's will; this leads to X's bad emotional state; X does not lose self-control, does not try to change the situation and continues to bear it'

In other words, the verb *терпеть* denotes an attitude of a person toward hardships and sufferings. For some reason, this person accepts a difficult condition, makes no attempt to interrupt this state by performing some action and remains in this state.

2.1.2 Aspectual properties

Терпеть 1.1 is an imperfectivum tantum. The delimitative derivative *потерпеть* is no aspectual correlate of *терпеть*; it means '*терпеть* for some time'. A doctor may say to a child as an unpleasant procedure starts: *Сейчас может быть больно, но ты немного* ***потерпи*** 'It may be a little painful; you'll have *to grin and bear the pain*'.

2.1.3 Constructions

Терпеть 1.1 may be used with or without an object (typical objects of *терпеть 1.1* refer to states related to experiencing physical or mental pain: *боль*, *муки*, *страдания*). In both cases, it may refer to a patient acceptance of a temporary pain or discomfort as well as a general attitude to life. The delimitative derivative *потерпеть* tends to be used without an object:

- *«**Потерпи**! – соседи хором говорят. –*
  Милого побои не долго болят!» (Nikolai Nekrasov, *Katerina*).

2.1.4 Derivatives

Verbs: *вытерпеть <боль>*, *стерпеть <обиду>*, *перетерпеть*:

- *Сейчас – только бы лечение как-нибудь **перетерпеть**!* 'If only he could somehow *get through* the treatment' (Alexander Solzhenitsyn, *Cancer Ward*)

Noun: *терпение*

Adjectives and corresponding adverbs: *терпеливый*, *нестерпимый*; *терпеливо*, *нестерпимо*:

- *И ему **нестерпимо** представилось, что ещё это всё он должен напрягаться делать, неизвестно зачем и для кого* 'It seemed *intolerable* to have to bring himself to do all these things, goodness knows who or what for' (Alexander Solzhenitsyn, *Cancer Ward*).

2.1.5 Cultural evaluation

The ability to endure temporary suffering, e. g. physical pain, and never complain is generally valued as a manifestation of courage and fortitude. On the other hand, this attitude is open to question: some people ask *Почему я обязан терпеть эти мучения?* 'Why am I obliged to endure these torments?' A physician may say, *Лучше не терпеть боль, а принять болеутоляющее* 'It is better not to bear pain but to take a pain-killer'.

The general attitude to life denoted by *терпеть 1.1* is valued in traditional Russian culture. Many Russian proverbs emphasize the importance of the ability to accept sufferings without protest: *Христос терпел и нам велел* 'Christ endured/suffered and ordered us to suffer'; *Не потерпев, не спасешься* 'Without suffering/enduring/bearing, one will not be saved'; *С бедой не перекоряйся, терпи!* 'Do not argue with misfortune, bear it', etc. Anna Gladkova noted that it is difficult to find proverbs with similar meanings among common English proverbs or sayings (Gladkova, 2007, p. 144).

On the contrary, Russian revolutionaries of the 1860s regarded *терпение* in this reading as something very bad. Murray B. Peppard observed that the poem "Katerina" (quoted above) was an example of Nikolai Nekrasov's "reworking of a folklore source": an old folk song tells how a young peasant submits to the will of her in-laws and prides herself on her patience; but Nekrasov converted the sense of the poem into an attack on such docile behavior (Peppard, 1967: 97). Consider also the following couplet *Чем хуже был бы твой удел, / Когда б ты менее терпел?* 'How much worse would be your lot in life / If you should cease your passive enduring?' (Nikolai Nekrasov, *On the Volga*).

The quality referred to with the adjective *терпеливый* (and the corresponding noun *терпеливость*) may get different evaluations as well. To quote Vladimir Solovyov, a famous Russian philosopher:

- *Терпеливость* (как добродетель) есть только страдательная сторона того душевного качества, которое в деятельном своем проявлении называется великодушием, или духовным мужеством. Тут почти вся разница исчерпывается субъективными оттенками, не допускающими твердых разграничений. <…> С другой стороны, единство внешних признаков может и здесь <…> прикрывать существенное различие этического содержания. Можно терпеливо переносить физические и душевные страдания или вследствие малой восприимчивости нервов, тупости ума и апатичности темперамента – и тогда это вовсе не добродетель; или вследствие внутренней силы духа, не уступающего внешним воздействиям,– и тогда это есть добродетель аскетическая (сводимая к нашей первой нравственной основе); или вследствие кротости и любви к ближнему (caritas), не желающей воздавать злом за зло и обидой за обиду,– и в таком случае это есть добродетель альтруистическая (сводимая ко второй основе: жалости, распространяемой здесь даже на врага и обидчика); или, наконец, терпеливость происходит из покорности высшей воле, от которой зависит все совершающееся,– и тогда это есть добродетель пиэтистическая, или религиозная (сводимая к третьей основе) (*The Justification of the Good*).

  Patience (as a virtue) is only the passive aspect of that quality of the soul which, in its active manifestation, is called magnanimity of spiritual fortitude. The difference is almost entirely subjective, and no hard and fast line can be drawn between the two. <…> On the other hand, the identity of the external expression may <…> conceal important differences in the moral content. A man may patiently endure physical or mental suffering owing to a low degree of nervous sensitivity, dullness of mind and an apathetic temperament, and in that case patience is not a virtue at all. Or patience may be due to the inner force of the spirit, which does not give way to external influences – and then it is an ascetic virtue (reducible to our first basis of morality) or it may arise from meekness and love of one's neighbors (*caritas*), which does not wish to pay back evil for evil and injury for injury – and in that case it is an altruistic virtue (reducible to the second principle – pity, which here extends even to enemies who inflict the injury). Finally, patience may spring from obedience to the higher will upon which all that happens depends – and then it is a religious virtue (reducible to the third principle) (Solovyof 1918: 103-104).

2.2 Терпеть 1.2

2.2.1 Explication

X терпит 'A person X wants something to happen; however s/he does not show his/her desire and does not act to precipitate it'

No "painful situation" is implied. The verb suggests rejecting a possible scenario of precipitating the desired event. A metonymical shift of this meaning is illustrated by the expressions *время терпит* 'time would wait; there's still time' and *время не терпит* 'time would not wait'.

2.2.2 Aspectual properties

Терпеть 1.2 as well as *терпеть 1.1* is an imperfectivum tantum. The delimitative derivative *потерпеть* is no aspectual correlate of *терпеть*; it means '*терпеть* for some time', e. g. ***потерпи**, и я все тебе отдам* 'be *patient*, and I will pay back everything'.

2.2.3 Constructions

Since no "painful situation" is implied, the verb *терпеть 1.2* is intransitive.

2.2.4 Derivatives

Some of the derivatives are morphologically the same as of the former meaning: the noun *терпение*, the adjective *терпеливый* (together with the adverb *терпеливо* and the noun *терпеливость*), the verb *вытерпеть* (in this meaning it is used with no object and for the most part with negation).

However, this lexical meaning has its own derivatives, among them a very peculiar noun *нетерпение* (consider the phrase *сгорать от нетерпения* 'to burn with impatience'). In general, the derivatives of the verb *терпеть* in this lexical meaning tend to include (or to collocate with) negation and to suggest that the person who wants something rejects the scenario of not showing his/her desire and not acting to precipitate the desired event: *нетерпеливый*, *нетерпеливо*, *не стерпеть*, *не утерпеть*, *не вытерпеть*, *не терпится*, *невтерпеж*: e. g. *не утерпел, чтобы не сказать, засмеяться*, etc. 'could not help saying, laughing, etc.'; *Баба тоже не стерпела – кочергой его огрела* 'The old woman did not endure it either and did not restrain herself from hitting him with a poker'.

Yet another meaning has evolved from the connotations of the meaning in question of the verb *терпеть*, namely, a reference to laborious task performed without expecting immediate result. This connotation has become a lexical meaning of the noun *терпение*. The adjective *терпеливый* may also be used in this way, e. g. *И Кларе открылся его затылок: как у мальчика слабый затылок, но обработанный **терпеливым** умелым парикмахером* 'Klara could see the back of his neck. It was scrawny, like that of a little boy, but it had received the *unhurried* attention of a skillful hairdresser' (Alexander Solzhenitsyn, *In the First Circle*). It should be noted that the verb itself does not have this lexical meaning; in other words, it never refers to performing laborious task.

2.2.5 Cultural evaluation

The ability to wait and to work without expecting immediate result is positively valued in the linguistic worldview. The following proverb refers to such a work: *Терпенье и труд все перетрут* 'Perseverance and labor will win'.

3 *Терпеть 2* (tolerance)

3.1.1 Explication

X *терпит* Y 'A person X does not like some other person or a thing Y; however, X expresses no dissatisfaction'

In the context of negation, a distinctly negative evaluation of Y by X is emphasized: *X не терпит / терпеть не может Y* 'X does not / cannot tolerate Y', e. g. *Я* ***не терплю*** *ресторанов, водочки, закусочек, музычки – и задушевных бесед* (Vladimir Nabokov, *Drugie Berega*; cf. the English version in his *Speak, Memory*: *I happen to have a morbid dislike for restaurants and cafés... I detest crowds, harried waiters, Bohemians, vermouth concoctions, coffee, zakuski, floor shows and so forth*); *Они сидели и беседовали как равный с равным, вполне приязненно, хотя каждый из них презирал и* ***терпеть не мог*** *другого* 'They sat and talked as equals, amicably, even though each despised and *loathed* the other' (Alexander Solzhenitsyn, *In the First Circle*).

3.1.2 Aspectual properties

The perfective quasi-correlate *потерпеть* in this meaning is almost exclusively used with negation, e. g. …*она знает толк в нарядах, в книгах и в хорошей обстановке, и у себя дома* ***не потерпела бы*** *такой комнатки, как эта, где пахнет сапогами и дешевою водкой* '…she has a good taste in dress, in furniture, in books, and in her own home she *would not have put up with* a room like this, smelling of boots and cheap vodka' (Anton Chekhov, *Neighbors*); *Я* ***не потерплю*** *в своем доме воров* 'I *cannot put up with* thieves in my house' (Anton Chekhov, *An Upheaval*).

3.1.3 Constructions

Терпеть 2 requires a direct object, which may only be omitted in case it is evident from the context, e. g. *Василисе казалось, что никто ее больше не замечает, никто с ней не считается, а только* ***терпят*** 'It seemed to Vasilisa that nobody noticed her anymore, nobody gave he consideration; they *put up with her* that was all' (Valentin Rasputin, *Vasili and Vasilisa*). (Consider, however, the famous Prayer of the Optina Elders mentioned below, in the *Concluding remarks*.)

3.1.4 Derivatives

The adjectives *терпимый* and *нетерпимый* (and, accordingly, the nouns *терпимость* and *нетерпимость*) are derived from this meaning of the verb *терпеть*. These words refer to the ability / inability of being patient and not hostile towards the views, opinions and behavior of other people (roughly, 'tolerant/intolerant'). They should be distinguished from the passive adjective *терпимый* and its negated form *нетерпимый* (roughly, 'tolerable/intolerable'). Consider the examples from the Russian National Corpus: …*будьте* ***терпимы*** *к нам, как и мы* ***терпимы*** *к вам* (Василий Аксенов); *Даже* ***терпимого*** *Павла Алексеевича он умел вывести из себя* (Людмила Улицкая); *Экстремальный жизненный опыт не сделал*

Райх-Раницкого ни более снисходительным к писательскому тщеславию, ни более ***терпимым*** *к людским слабостям вообще* (Ревекка Фрумкина), on the one hand (‘tolerant’), and *«Ну как, – говорю я, – что там передают насчёт погоды? Ветер с востока?» «Нет, – радостно отвечают москвичи, – ветер юго-западный до умеренного». – «Ну, если до умеренного, – говорю, – это ещё* ***терпимо****»* (Фазиль Искандер); *А тут круговая порука. Разве это в советской школе* ***терпимо****?* (Юрий Домбровский); *Он недоумевал: «Как же можно так жить? Вы же все нищие!» Соседи, в свою очередь удивлённые его восприятием того, что им казалось вполне* ***терпимым****, обычным, отвечали: «Почему же? У нас есть всё необходимое – жильё, еда, мы одеты, работаем, живём дружно…»* (Ирина Архипова); *…это было ещё* ***терпимое*** *неудобство* (Вячеслав Пьецух), on the other (‘tolerable’).

3.1.5 Cultural evaluation

The attitude referred to with the verb *терпеть 2* does not have a component of evaluation since the reason why the person chooses not to express his/her dissatisfaction may lie in selfish considerations. It may also be perceived as a kind of humiliation by its object.

The same is true with regard to *терпимость* and *нетерпимость*. To quote Vladimir Solovyov again:

- Особая разновидность терпеливости есть качество, которому присвоено по-русски неправильное в грамматическом отношении название терпимости (passivum pro activo). Так называется допущение чужой свободы, хотя бы предполагалось, что она ведет к теоретическим и практическим заблуждениям. И это свойство и отношение не есть само по себе ни добродетель, ни порок, а может быть в различных случаях тем или другим, смотря по предмету (наприм., торжествующее злодеяние сильного над слабым не должно быть терпимо, и потому «терпимость» к нему не добродетельна, а безнравственна), главным же образом – смотря по внутренним мотивам, каковыми могут быть здесь великодушие, и малодушие, и уважение к правам других, и пренебрежение к их благу, и глубокая уверенность в побеждающей силе высшей истины, и равнодушие к этой истине (*The Justification of the Good*).

  A particular variety of patience is the quality which is designated in the Russian language by the grammatically incorrect term ‘*terpimost’* – tolerance (*passivum pro activo*). It means the admission of other people’s freedom even when it seems to lead to error. This attitude is in itself neither a vice nor a virtue, but may in different circumstances, become either. It depends on the object to which it refers (thus injury of the weak by the strong must not be tolerated, and ‘tolerance’ of it is immoral and not virtuous), and still more, on the inner motives from which it arises. It may spring from the magnanimity or from cowardice, from respect for the rights of others and from contempt of the good of others, from profound faith in the conquering power of the higher truth and from indifference to that truth (Solovyof 1918: 104).

However, more often than not *терпимость* is evaluated positively. It is often named among other human virtues. Anna Gladkova has suggested that it is regarded in the Russian “naïve” axiology as an essential quality in interpersonal relations, especially within a family. She adds

that this quality is more often associated with women than with men. Many women providing information about themselves on a Russian website of single people looking for partners write that they are *терпимы* (along with such qualities as being *оптимистичны*, *жизнерадостны*, *общительны*) presenting it as one of their virtues. Apparently, they do so because they think that this quality will be attractive to men and that it will convey their ability to accept weakness and faults in men, e. g. excessive drinking (Gladkova, 2007, p. 151).

In her contrastive analysis of two virtues in English and Russian (*tolerant* and *терпимый*), Anna Gladkova states that *tolerant* has a more "social" character since it is an attitude towards something seen as deviating from social norms. *Терпимый* is more personal since it is a reaction towards personal offense. She attributes the differences in meaning to different cultural ideas prevailing in Russian- and English-speaking society. *Tolerant* is related to the idea of 'not imposing one's views on others' and to the idea of social harmony as an opportunity for people to behave and think in the way they want. *Терпимый* is about not developing negative reactions to other people's behavior and about maintaining the social harmony of positive feelings among people (Gladkova, 2007, p. 159-164). However, *терпимость* may be used to refer to the respect for other people's views. Consider the following reasoning: *Терпимость – любимая категория и высшая ценность Самойлова. На переходе к терпимому обществу мы должны прежде всего научиться уважать любое другое мнение, даже не нравящееся нам. / Дай-то Бог. Всем нам* 'Tolerance is Samoilov's favorite category and supreme value. At the transition to a tolerant society, we have to learn to respect other people's opinion even when we dislike it. / God grant it. To all of us' (Alexander Solzhenitsyn on David Samoilov).

4 *Терпеть 3* (lexical function)

4.1.1 Explication

X терпит Y 'Something bad Y happens to X'

The verb is used in collocation to refer to experiencing a negative condition, e. g. *терпеть поражение* 'to suffer defeat'. Since it is a manifestation of the lexical function OPER, it is semantically void in the context of its keywords; the semantic component 'something bad' is included in the meaning of the possible keywords. It is idiomatic to a certain extent: it only collocates with nouns denoting 'bad conditions'; however, all nouns denoting 'bad conditions' do not collocate with it.

4.1.2 Aspectual properties

If the keyword denotes a state, *терпеть 3* is an imperfectivum tantum (and no delimitative verb can be derived from it: consider *терпеть нужду*, but not **потерпеть нужду*). If the keyword denotes an event, he may be used as a "trivial" aspectual correlate of *потерпеть* (Зализняк, Шмелев, 2000, p. 56): *потерпеть/терпеть поражение*; *потерпеть/терпеть неудачу* 'to suffer a setback'. A special type of aspectual correlation is represented by such constructions as *потерпеть / терпеть аварию* 'to get into accident'; *бедствие*,

катастрофу ‘to suffer a disaster’; *кораблекрушение* ‘to suffer a shipwreck’. The imperfective member of such a pair can denote a state resulting from the event denoted by the perfective member, this state may last for some time.

4.1.3 Constructions

As a manifestation of the LF OPER, *терпеть 3* requires a direct object (a keyword). It is rare for it to be used with no object when “the bad condition” can be inferred from the context (e. g. an indirect object). Consider: *…такой просвещенный гость, и* ***терпит****, от кого же? от каких-нибудь негодных клопов* ‘Such a cultured visitor and *to suffer* from what? – Worthless bugs’ (Nikolai Gogol, *The Inspector General*).

4.1.4 Derivatives

Терпеть 3 has few derivatives. The substantivized participle *потерпевший / потерпевшая* ‘victim; injured party’ is a legal term. The syntactic valence of an object is reduced, but the semantic valence remains obligatory and is to be filled on the basis of the situation. In addition, *терпеть 3* has the verbs *претерпеть* and *претерпевать* (a “potential” aspectual pair in terms of (Зализняк, Шмелев, 2000)) amond its derivatives as well as the saturative verb *натерпеться*.

4.1.5 Cultural evaluation

Since the situation denoted by the collocation with *терпеть 3* is beyond control, it has no evaluation in the Russian “naïve” axiology; meanwhile the disastrous state of the person in question can arouse sympathy.

5 Concluding remarks

The verb *терпеть* as it usually is with polysemous words (Шмелев 1973) has vague uses, which are difficult to assign to one of its lexical meanings. Thus, the famous Prayer of the Optina Elders addresses God with the following words:

- *…научи меня молиться, надеяться, верить, любить,* ***терпеть*** *и прощать!*

The attitude denoted by *терпеть* is mentioned along with other important virtues. One might think that we deal with a very rare absolutive usage *терпеть 2*, which roughly means ‘to be tolerant’. On the other hand, (Gladkova, 2007, p. 144) cites this prayer and translates the above words as “…teach me to pray, to hope, to believe, to love, to endure and to forgive” implicitly suggesting that a reference is made to the ability of accepting sufferings (*to endure*). It should be noted that in the accepted English version of the prayer, the verb *to suffer* is used: *teach me to pray, to believe, to hope, to suffer, to forgive, and to love*. The more so, the verb *to suffer* is somewhat vague and has tolerating other people’s defects among its readings: e. g. the Church Slavonic *долготерпеливый* is rendered into English as *long-suffering*. The morale of this story is that the semantic potential of *терпеть* is by no means unique to Russian; however, the linguistic characteristics of the verb (including its derivational patterns) are

language-specific in the sense that the whole set of its lexical meanings (as well as of its derivatives) reflects Russian cultural values.

Bibliography

Gladkova, Anna. 2007 *Russian Emotions, Attitudes and Values: Selected topics in cultural semantics*. A thesis submitted for the degree of Doctor of Philosophy of the Australian National University. Canberra.

Peppard, Murray B. 1967. *Nikolai Nekrasov*, New York: Twayne Publishers, Inc.

Solovyof, Vladimir. 1918. *The justification of the good: an essay on moral philosophy*, translated from Russian by Nathalie A. Duddington. London: Constable and Company Ltd.

Зализняк, Анна А. & А.Д. Шмелев 2000. *Введение в русскую аспектологию*. Москва: «Языки русской культуры», 2000.

Шмелев, Д.Н. 1973. *Проблемы семантического анализа лексики (на материале русского языка)*. Москва.

Шмелев, А.Д. 2003. Терпимость в русской языковой картине мира. In *Философские и лингвокультурологические проблемы толерантности*, 111-125. Екатеринбург.

Online Tutoring and Collocations

Serge Verlinde

Leuven Language Institute
Katholieke Universiteit Leuven

Abstract

At our institute, we are currently developing an online tutoring tool for French, consisting of a spelling, grammar and lexical checker. This application integrates information usually found in dictionaries and grammars that can adjust itself to the user's input and needs like other spelling and grammar checkers. Unlike these checkers, however, this tool does not necessarily correct every mistake found in the text, because automatic correction is too difficult for many items. It essentially identifies words or patterns which could contain errors. It is then up to the student, assisted by very specific contextual feedback, to reread his text and spot errors.

The application also includes a lexical checker, more specifically aimed at collocations. On the one hand, students often fail to use appropriate collocations because they are unfamiliar with typical word combinations. On the other hand, they often use incorrect collocations (ex. *demander une question*, *regagner des forces*, *une entreprise profitable*).

The collocation tool suggests relevant combinations for nouns, classified according to a 'light' version of Mel'čuk's lexical functions. It also scans submitted texts to find odd combinations like those mentioned above and to suggest alternatives.

An online collocation dictionary of Spanish

Orsolya Vincze (1), Estela Mosqueira (1) and Margarita Alonso Ramos (1)

(1) Facultade de Filoloxía, Universidade da Coruña
Campus da Zapateira, s/n, 15071 Coruña (Spain)
ovincze@udc.es|estela.mosqueira@udc.es|lxalonso@udc.es

Abstract

DiCE is an online dictionary of Spanish collocations, which provides semantic and combinatorial information on lexical units. It makes use of the typology of lexical functions (Mel'čuk et al. 1995), together with natural language glosses to describe the semantic content of collocates. We offer a general presentation of the content of the dictionary, followed by a description of the user's interface and a brief insight into the lexicographer's interface. The main feature of DiCE is that it is conceived as a part of a language-learning environment which combines dictionary, corpus and teaching materials.

Keywords

collocations, lexical functions, online dictionary, learning tools, Spanish as a second language

1 Introduction

In this paper, we present the *Diccionario de Colocaciones del Español* (DiCE), a web-based collocation dictionary of Spanish, which is available online since 2004 with its database being constantly modified and expanded. Although the dictionary has been presented on various occasions (e.g. Alonso Ramos, 2005, 2006, 2008, 2010), hereby, we will focus on some features that have not been described in previous publications.

Similarly to Dicouèbe (Polguère, 2000, Jousse & Polguère, 2005) and DicoInfo (L'Homme, 2009), DiCE constitutes an online implementation of the principles of lexical description introduced in the *Explanatory Combinatorial Lexicology* (Mel'čuk et al., 1995). In addition to providing a theoretically-based description of collocations, DiCE aims to be a useful tool not only for researchers, but also for the general public, that is, native and non-native speakers of Spanish. This is achieved, on the one hand, by adapting the information offered by the dictionary to the needs of general users, for instance, by paraphrasing *lexical functions* with natural language glosses. On the other hand, the web interface is designed to enable flexible

access to the electronic lexical database, satisfying users ranging from researchers through language learners to lexicographers working on DiCE.

In the following sections, we offer a general presentation of the content of the dictionary, followed by a description of the user's interface and a brief insight into the lexicographer's interface.

2 General presentation – the content of DiCE

DiCE is essentially a collocation dictionary: its main objective is to describe restricted lexical co-occurrence, although it also deals with semantic derivatives. More precisely, it concentrates on both paradigmatic and syntagmatic lexical relations controlled by a lexical unit (LU). For now, the list of lemmas treated in the dictionary is limited to the semantic field of emotions, so that DiCE specifies approximately 19500 lexical relations (values of lexical functions).

The lexicographical description of each LU can be divided into semantic and combinatorial information. As for the semantic information, the entry of each LU provides a) a *semantic tag* that represents the generic meaning; b) the *actantial structure* representing the participants of the situation designated by the noun; c) corpus examples, most often derived from the online *Corpus of the Real Academia Española* (CREA); and d) quasi-synonyms (QSyn) and quasi-antonyms (QAnti) of the LU.

The combinatorial information offered by the dictionary is of two types: syntactic and lexical combinatorics. The syntactic combinatory information of the LU is shown in the Government Pattern (*esquema de régimen)* section, where we specify the projection of the semantic valency structure of the LU onto its syntactic valency structure and, in addition, the subcategorization information associated with the latter. To illustrate this, as shown in Figure 1, in the case of ALEGRÍA 'joy' whose propositional form is *alegría de individuo X por hecho Y* 'person X's joy over fact Y' we specify that, for instance, semantic actant X can be realized as a prepositional phrase headed by the preposition *de,* as a possessive determiner or as an adjective (see examples 3, 10 and 4 respectively). The lexical combinatory information is displayed in the section Collocations. In what follows, we focus on lexical combinatorics.

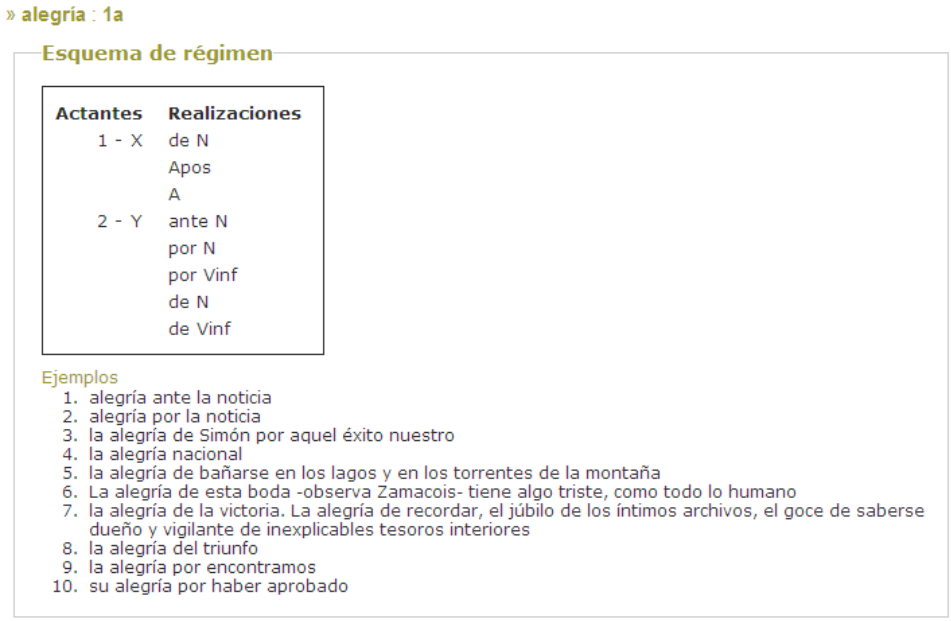


Figure 1: Syntactic combinatorial information on UL *alegría 1a*

Taking a specific LU as the starting point, the user can choose between five different groups of lexical correlates:

1. *Attributes of the participants*: Under this heading, we have grouped those attributes or nouns that refer to the participants of the situation designated by the LU. For example, in the entry for ALEGRÍA 'joy', the user finds *loco de alegría* 'crazy with joy' or *exultante* 'joyful, exultant', both referring to the participant who is feeling a lot of joy;

2. *LU + adjective*. Here, the user finds adjectives that co-occur with the LU;

3. *Verb + LU*: In this section, we have grouped the verbs that take the LU as a direct complement or as a prepositional complement, *e.g. provocar dolor* '[to] cause pain';

4. *LU + verb*: This section contains verbs that take the LU as the grammatical subject, *e.g. la alegría se desvanece* 'joy disappears';

5. *Noun + de LU*: Here, we find noun collocates that precede the LU, introduced by the preposition *de* 'of'; *e.g. arrebato de celos* 'a fit of jealousy'.

Once a user has entered one of these sections, they will find a list of collocates or semantic derivatives preceded by an LF, and followed by a gloss and one or more examples. In the gloss we intend to give a brief indication of the meaning of the collocate in relation to the base. So, the gloss *intenso* 'intense' serves to group various adjectives such as *desbordante* 'overwhelming', *enorme* 'enormous', and *indescriptible* 'indescribable', which, in combination with the noun ALEGRÍA 'joy', fulfill the same role, although they do not have strictly the same meaning. Using glosses to describe the meaning of collocations proved to be a very useful feature especially for learners, who may have a problem interpreting or choosing collocations without explicit information on their meaning, generally missing from

conventional collocation dictionaries. For instance, the meanings of the following adjectives used with the noun ODIO 'hatred' are described in the glosses as follows:

(1) *mortal* 'lethal', glossed as *intenso* 'intense' [Magn(*odio*)]

(2) *declarado* 'declared', glossed as *que se manifiesta* 'manifest' [A_1Manif(*odio*)]

(3) *eterno* 'eternal', glossed as *que dura mucho* 'long-lasting' [$Magn_{\_temp}$(*odio*)]

(4) *larvado* 'latent', glossed as *que no se manifiesta* 'that doesn't manifest itself'[A_1nonManif(*odio*)]

3 The user's interface

The user's interface consists of three main components: 1) the dictionary itself, 2) the advanced search component, and 3) the didactic module.

3.1 The dictionary component

This component allows the user to access the dictionary in the more traditional way, through the list of lemmas. Each lemma is associated with a list of lexical units (LUs), where corresponding semantic and combinatorial information can be found (see above).

3.2 The advanced search component

The *Consultas avanzadas* 'advanced search' component serves to carry out specific queries. Rather than looking up the list of collocates of a given LU, it helps us find the answer for specific questions.

The user is provided with four types of search tools: 1) *direct search*, 2) *inverse search*, 3) *what does it mean?*, and 4) *writing aid*.

3.2.1 Direct search

Consultas directas 'direct search' allows the user to find collocations described by a given LF. As an answer to the query for combination Magn+$Caus_2Oper_1$, the system returns all collocations described by this sequence of LFs stored in the database. To restrict this search, the user has a further option of specifying the lemma of the base and its LU. For instance, as shown in Figure 2, one can restrict the search for collocates described by the LF Magn+$Caus_2Oper_1$ specifying the lemma ALEGRÍA and choosing one of its LUs, in this case *1a*. Note that, in order to facilitate the choice of a specific LU, in each case, an example is visualized together with the numeration used to code LUs in the dictionary entry.

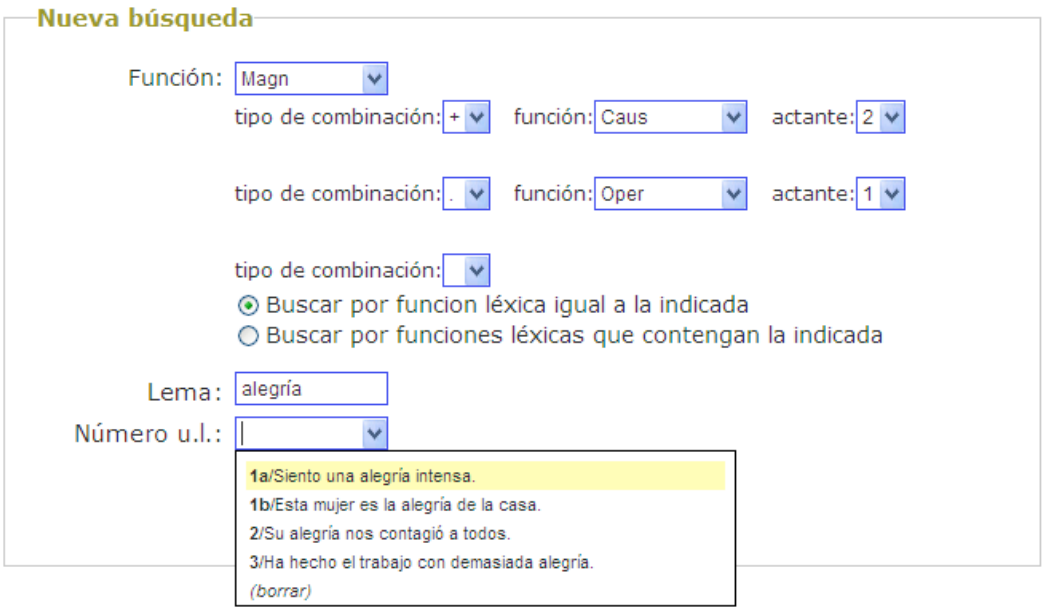


Figure 2: Direct search for *Magn+Caus$_2$Oper$_1$(alegría 1a)*

3.2.2 Inverse search

The option *Consultas inversas* 'inverse search' allows the user to find the base of a collocation starting from the collocate. After having indicated the collocate, the search can be further restricted by specifying the LF associated with the collocation. Figure 3 shows a sample of the results obtained from the search for the collocate *guardar* 'keep'. Through this query the user can learn that the same collocate, in this case the verb *guardar* can be combined with different bases in order to constitute collocations that are codified by different LFs. For instance, *guardar rencor* 'bear a grudge' is described by the LF ContOper$_1$ while *guardar sorpresa* 'have a surprise in store (for sb)' is codified by LF CausFunc$_2$.

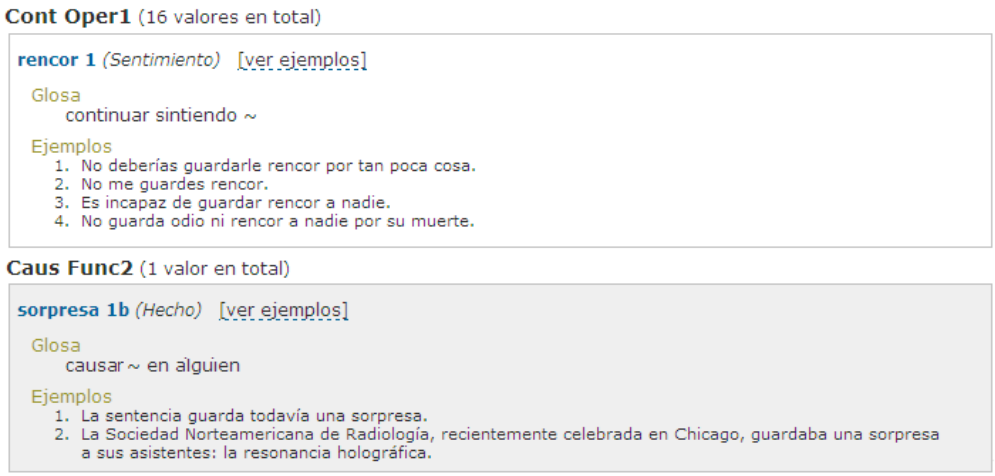


Figure 3: A sample of the results of an inverse search for *guardar* as a collocate

3.2.3 What does it mean?

The module *¿Qué significa?* 'What does it mean?' is oriented towards comprehension. It serves to find which LF – and gloss – codifies the relation between a given base and a collocate. For example, as shown in Figure 4, we can learn that *ligero* expresses the meaning 'small in degree' when combined with the base *arrepentimiento 1* 'repentance'.

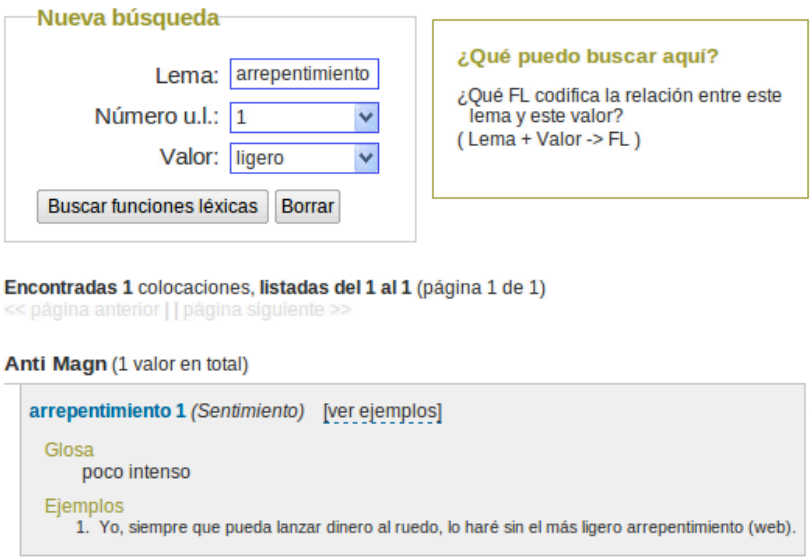


Figure 4: Results of a search for the collocation *arrepentimiento ligero* with the *¿Qué significa?* tool

3.2.4 Writing aid

The option *Ayuda a la redacción* 'writing aid' is intended to resolve questions concerning lexical combinatorics raised by any speaker of Spanish, including learners and native speakers. At this moment, we offer the following two types of aid:

1. The first kind of aid allows the user to check whether a given base can co-occur with a given collocate. In Figure 5, we show the result of a query for the collocation *arrepentimiento ligero* 'light repentance'.

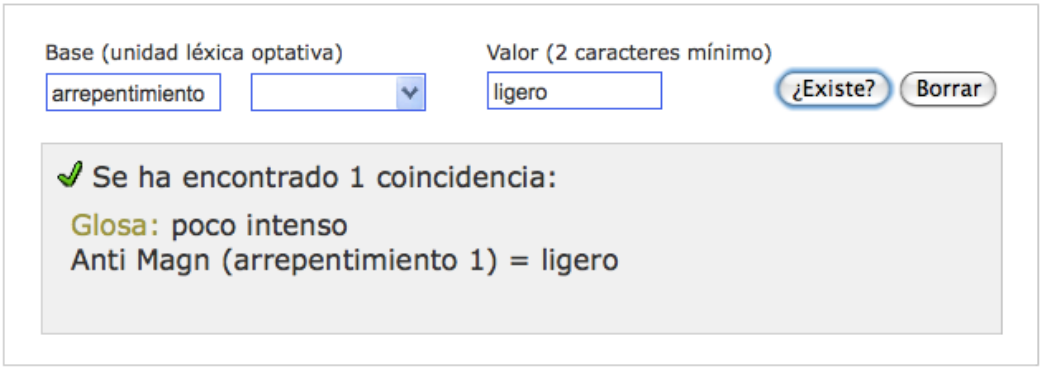


Figure 5: Checking the collocation *arrepentimiento ligero* with the Writing aid tool

2. The second aid enables users to find collocates corresponding to a specific meaning, codified by a gloss, and a syntactic scheme (under "tipo"). Figure 6 shows a search for a collocate adjective of *amor I.1a* 'love' with the meaning 'felt for one another', for which the tool returns the collocation *amor correspondido.*

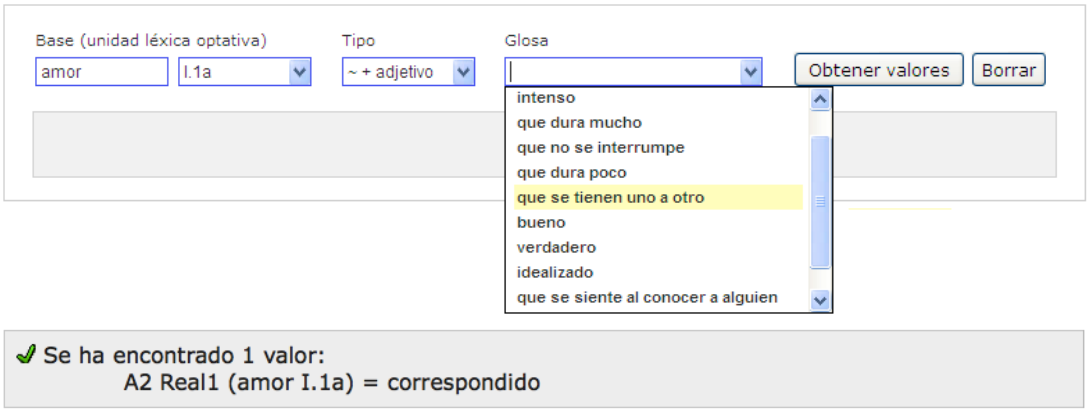


Figure 6: Finding collocates of *amor* I.1a with the meaning 'felt for one another' using the Writing aid tool

3.3 The didactic module

The aim of the exercise module, still in development, is to provide the user with learning material concentrating on collocations. For now, it is limited to a few sections containing exercises related to a particular topic, among others, an introduction to the use of DiCE itself.

Our mid-term goal is to exploit the DiCE database integrating the dictionary with a more complete didactic module, providing an online language learning environment. For further support of the learner, we are planning to offer users the option of creating their own learning space in which they can administrate personal collocation lists, annotations, performance scores and problems identified with respect to specific collocations or collocation types. Furthermore, the DiCE forms part of the COLOCATE Project where we intend to integrate the lexical database with a corpus interface and a checker tool that will provide aid for users on collocations encountered in reading as well as writing tasks[1].

4 The lexicographer's interface

The lexicographer's interface allows the editors of DiCE to carry out instant modifications in the lexical database via the web. Essentially, there are two ways of editing the dictionary: either through viewing the *User's Interface* or through the *Administration Area*.

[1] The COLOCATE Project is being conducted with collaboration of researchers from the Universitat Pompeu Fabra and the Universidade da Coruña. Experiments carried out by Wanner and his colleagues have shown promising results. See Wanner et al. (this volume).

4.1 Editing DiCE through the User's Interface

Editors have the option of modifying the DiCE database directly from the User's Interface view. This allows quick corrections and modifications of the content while browsing the dictionary. For instance, a lexicographer can access the edition area to correct a mistake in the description of a concrete collocation by simply clicking on the corresponding icon, see Figure 7.

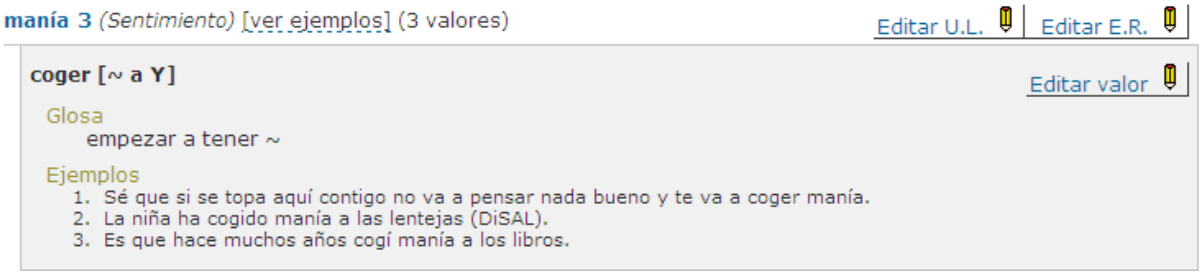


Figure 7: Lexical entry of the collocation *coger manía* 'take a dislike (to sb)' with the "Edit" icons visible in the upper right-hand corner.

4.2 Editing DiCE through the Administration Area

Lexicographers are provided with a more efficient tool to carry out modifications, remove or hide existing data and incorporate new information in the database through the two main options of the Administration Area: 1) *Modification of lexical information* and 2) *Mass update*.

4.2.1 Modification of lexical information

In this section, the lexicographer has access to the database through the list of lemmas to edit both semantic and combinatory information of LUs, such as the semantic tag, the actantial structure and the government pattern of LUs, EuroWordNet IDs[2] and the descriptions of collocations themselves. In Figure 8 we show the lexicographer's interface screen for the lemma AMISTAD.

[2] We provide the ID assigned in Spanish EuroWordNet for each LU. This information comes from the research carried out during a project focusing on linking DiCE with the Spanish EuroWordNet) (see Wanner et al. 2004).

Unidad léxica			EWNET	Esquema de régimen	Colocaciones	Operaciones		
Número de U.L.	Forma proposicional.	Etiqueta semántica	EWNET	Esquema de régimen	Colocaciones	Publicado	Editar	Eliminar
1	amistad de individuo X hacia individuo Y	Sentimiento						
	Ejemplos : 1 - No son novios, entre ellos sólo existe una buena amistad (Clave) 2 - En Sahagún contaba con la amistad y la hospitalidad de Martín y de Zulema							
2a	[individuo Y] es una amistad de individuo X	Individuo						
	Ejemplos : 1 - Me presentó a una amistad de la infancia (larousse) 2 - Tengo algunas amistades en Francia (DUE) 3 - felicitaba el año nuevo a sus amistades							
2b	[individuos Y] son las amistades de individuo X	Individuo						
	Ejemplos : 1 - Tiene amistades en el ministerio que lo apoyarán (DiSAL) 2 - ese negocio se lo debo a las amistades (Lema) 3 - se valió de todas sus amistades para poder concretar contactos en la pequeña comunidad del cine internacional							

Figure 8: Lexicographer's interface screen for editing information on LUs of the lemma AMISTAD

4.2.2 Mass update

This option allows editors of the dictionary to carry out mass modifications of glosses, government patterns, collocate lemmas and lexical functions. So far, editors have found this option especially useful in adding or modifying glosses of large groups of collocations. As we show in Figure 9, glosses of all collocations described by the LF $Oper_1$ and belonging to LUs with the semantic tag 'sentimiento' ('feeling') can be easily changed using this tool.

Actualizar la glosa a...

Nuevo valor de Glosa:

sentir ~

Actualizar Limpiar

650 colocaciones, **listadas de la 1 a la 20** (página 1 de 33)

<< página anterior | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | página siguiente >>

Lema	Unidad léxica	Función léxica	Glosa	Valor	Régimen
abatimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sentir	
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	tener	[~]
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sentir	[~]
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	experimentar	[~]
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	nutrir	[ART ~]
aburrimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sentir	[~]
aburrimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sufrir	[de ~]

Figure 8: Mass update of natural language glosses of collocations with the LF $Oper_1$

Similarly, it is possible to edit LFs of a large amount of collocations sharing the same collocate or base (for example, to change the combination of LFs that encodes collocations

with the collocate *enfermizo* 'unhealthy' from Magn+AntiBon to Magn+AntiVer), or to change the government pattern associated with collocations described by a given LF, etc. All these options are quite useful in making necessary changes or correcting errors in the database in a largely efficient way.

4.3 Exploiting DiCE as a corpus

We think that the examples found in the lexical entries of the dictionary can constitute valuable data for research. This is why we have included a tool to extract corpus examples from the dictionary in the Administration Area. Introducing a specific LF or combination and, optionally, a lemma, the lexicographer can download a .txt file with the corresponding example sentences. For this reason, we can say that the DiCE contains a corpus of collocations (Alonso Ramos, 2009). To illustrate this point, all examples containing collocations described by the LF $Func_1$ constitute a corpus of 3814 words; those that contain collocations described by $Oper_2$ amount to 3236 words; and, those illustrating cases of $IncepOper_1$, contain 3887 words, and so on. Collocation corpora obtained from the DiCE can be tagged and parsed in order to obtain a collocation Treebank for further investigation needs.

5 Conclusion and future work

A dictionary, as we conceive of it, is necessarily a project that is constantly in the course of development and indefinitely undergoing changes. In this way, the DiCE already has a version 1.0 (which was available from 2004) and the current version undergoing constant modifications since last year. It is changing not only in its content, but also in its interface and ways of access to information.

With respect to the content of the DiCE, future changes concern information on the frequency of use of collocations. After a long process of semantic disambiguation of corpus samples, we managed to assign frequency scores to the LUs contained in the dictionary, that is, the bases of collocations. As the next step, we will proceed to assign frequency information to collocations as a whole. Another piece of information we aim at adding shortly concerns assigning collocations to particular levels of L2 Spanish: this means specifying which collocations should be taught to learners of different levels (elementary, intermediate and advanced), with a view to extracting leveled teaching materials. With respect to the access to information, our goal is to develop a semantic typology of LFs (similar to the one proposed by Jousse, et al. 2008) that would allow the user to look up collocations with a semantic focus. For instance, if a user is searching for how to verbalize the meaning related to the phase of the starting fear, it would be convenient to find verb+object collocacions like *coger miedo* 'take fear of sg' as well as subject+verb collocations like *entrarle miedo* 'fear enters sb', *asaltarle miedo* 'fear assaults sb', or *invadirle el miedo* 'fear invades sb'. At this moment, these cannot be found in one single search, given that collocations are currently classified according to their syntactic structure.

As we have shown, the electronic format of DiCE and the codification of collocations through LFs and glosses turn out to be a clear advantage over conventional collocation dictionaries, but this is not enough. In line with the proposals put forward by Verlinde et al. (2009), the concept of dictionary is changing towards being a more flexible and more dynamic tool, which is more oriented to the users' needs; a tool that should be considered as a *leximat* (Tarp,

2008). Jousse et al. (2008) also prefer referring to this new concept as a *lexical site*, instead of a *dictionary*, due to the connotations of a linear vision carried by this latter term, while the first one proves to be a better model of lexical knowledge, as a constantly evolving network. Independently of the term we use to refer to these new lexical resources, the fact is that they have ceased to be stand-alone products, and they are necessarily integrated with other resources such as corpus and other dictionaries and glossaries. This is exactly the course of evolution we intend DiCE to take within the framework of the COLOCATE Project (see above).

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation and the FEDER Funds of the European Commission under the contract number FFI2008-06479-C02-01. We would also like to thank the anonymous reviewers for their valuable remarks and comments.

Bibliography

Alonso Ramos, M. 2005. Semantic Description of Collocations in a Lexical Database. In Kiefer, F. et al. (eds.). *Papers in Computational Lexicography COMPLEX 2005*, 17-27. Budapest: Linguistics Institute and Hungarian Academy of Sciences.

Alonso Ramos, M. 2006. Towards a Dynamic Way to Learn Collocations in a Second Language. In Corino, E., C. Marello & C. Onesti (eds.) *Proceedings of the Twelfth EURALEX International Congress*, 909-923. Torino: Accademia della Crusca, Università di Torino, Edizioni dell'Orso Alessandria.

Alonso Ramos, M. 2008. Papel de los diccionarios de colocaciones en la enseñanza de español como L2. In Bernal, E. & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*, 1215-1230. Barcelona: IULA, Documenta Universitaria.

Alonso Ramos, M. 2009. Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario? In Cantos Gómez, P. & A. Sánchez Pérez (eds.). *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*, 1191-1207. Murcia: AELINCO.

Alonso_Ramos, M. 2010. No importa si la llamas o no colocación, descríbela. Mellado, C. et al. (eds.), *Nuevas perspectivas de la fraseología del siglo XXI*, 55-80. Berlin: Frank & Timme.

Jousse, A. L. & A. Polguère. 2005. *Le DiCo et sa versión DiCouèbe. Document descriptif et manuel d'utilisation*. Versión du rapport 1.0 – 19 avril 2005, Montréal: Observatoire de linguistique Sens-Texte (OLST).

Jousse, A. L., A. Polguère & O. Tremblay. 2008. Du dictionnaire au site lexical pour l'enseignement/apprentissage du vocabulaire. In Grossmann, F. & S. Plane (eds). *Les apprentissages lexicaux. Lexique et production verbale*, 141–157. Villeneuve d'Ascq: Presses univeritaires du Septentrion.

L'Homme, M. C. 2009. *DiCoInfo: Le dictionnaire fondamental de l'informatique et l'Internet. Document descriptif et manuel d'utilisation*, Montréal: Observatoire de linguistique Sens-Texte (OLST).

Mel'čuk, I., A. Clas & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire,* Louvain-la-Neuve: Duculot.

Polguère, A. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In Heid, U., S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the Ninth EURALEX International Congress*, 517-527. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Tübingen: Niemeyer.

Verlinde, S., P. Leroyer & J. Binon. 2009. Search and you will Find. From Stand-alone Lexicographic Tools to User Driven Task and Problem-Oriented Multifunctional Leximats, *International Journal of Lexicography*, 23(1):1–17.

Wanner, L., M. Alonso & M. A. Martí. 2004. Enriching the Spanish *EuroWordNet* by *Collocations*. In *Proceedings of LREC* 2004, Vol. 4, 1087-1091, Lisbon: ELRA.

Wanner, L, G. Ferraro & R. Nazar (this volume). Collocations: A Challenge in Computer Assisted Language Learning. *Proceedings of the Fifth International Conference on Meaning-Text Theory* (MTT '11).

French-Spanish-Russian Pragmatemes Dictionary

Maria VOROBEY

(1) Laboratori fLexSem – Departament de Filologia Francesa i Romànica –
Universitat Autònoma de Barcelona –
08193 Bellaterra (Barcelona) –SPAIN
xavier.blanco@uab.cat

(2) Centre Tesnière – UFR SLHS – 30 rue Mégevand – 25030 Besançon cedex – FRANCE
sylviane.cardey@univ-fcomte.fr

Abstract

This paper relates, firstly, about pragmatemes, expressions restricted by the extralinguistic situation. Secondly, it deals with a contrastive approach to the study and shows semantic, pragmatic and cultural peculiarities of Russian, French and Spanish pragmatemes. Pragmatemes are very important in everyday life: we use them as greetings and civilities; we read them on signs, posters and packages. Therefore these stereotypical utterances are required at initial levels of foreign language acquisition. We propose a dictionary of pragmatemes in French, Spanish and Russian for beginner and intermediate students. All expressions are divided into 14 thematic groups. The microstructure of the dictionary contains the information about the conceptual representation, the situation in which a pragmateme occurs, its variants and synonyms.

Keywords

Phraseme; pragmateme; Meaning-Text theory; contrastive studies; teaching and acquisition of foreign languages.

1 Introduction

Push/Pull, No parking, Bon appetit! Best before and etc. We use these expressions in everyday life. We pronounce them as greetings and civilities; we read them on signs, posters and packages.

They are pragmatemes. According to Mel'čuk people do not speak most of the time in words but in phrasemes (Mel'čuk *et al.*, 1995). Unlike the case of semantic phrasemes (idioms, clichés, collocations), pragmatic phrasemes (pragmatemes) do not attract particular attention of lexicographers. However their study and description could be useful in applied linguistics,

specifically in foreign language teaching. The application of a contrastive analysis of pragmatemes will provide empiric results and reveal new aspect of the phenomenon. We propose a dictionary of pragmatemes in French, Spanish and Russian for learners of initial levels of foreign language acquisition.

2 Notion of "pragmateme"

I. Mel'čuk was the first to introduce the term of "pragmateme" and to give a fundamental approach to their study (Mel'čuk, 1998).

With regards to the notion of "pragmateme", we proceeded from the definition proposed by I. Mel'čuk who considers a pragmateme as "a compositional phraseme, if it is restricted in its signified and its signifier by the extralinguistic situation in which this phraseme is used" (Mel'čuk, to appear). We extend this definition.

By the term "pragmateme" we mean not only compositional phrasemes, but also a linguistic sign pragmatically bound by a its situation of use. Thus, a pragmateme can be represented by a single lexical unit, a phrase, or a sentence:

- a lexical unit: Sp. *¡Jesus!* [to someone sneezing];
- a phrase: Rus. *Парковка запрещена* (Parking prohibited[1]) [on a road sign];
- a sentence: Fr. *Pour votre sécurité, ce lieu est sous video surveillance* (For your safety, these premises are under video surveillance) [dans un espace public].

We consider this broad definition justified because in different languages both a single lexical unit and a multilexemic expression can correspond to the same conceptual representation in the same situation. Cf.: Fr. *Silence* and Rus. *Соблюдайте тишину* (Respect the silence) [on a sign in the library].

2. 1 Synonyms of pragmatemes

The extralinguistic situation determines the choice of an appropriate meaning or, sometimes, several meanings.

For example, in public places, in libraries, museums, theatres, there is a sign prohibiting the usage of mobile phones. The prohibition can be expressed differently: Sp. *Mantengan desconectados los teléfonos móviles* (Keep your mobile phones disconnected) and *Prohibido el uso del teléfono móvil dentro del recinto* (The use of mobile phones is prohibited on the premises) [on a sign].

[1] Throughout the work we will use brackets to provide the literal translation of the given pragmatemes in English.

In France, the access to the premises reserved for staff in restaurants or hotels is restricted by signs: Fr. *Interdit au public* (Prohibited to public) and *Entrée de service* (Service entry) [on a door]. These utterances are interchangeable and can appear both in the situation in question.

Thus, for the same extralinguistic situation and the same conceptual representation, two (or more) possible utterances exist. In this case we deal with *synonyms* of pragmatemes.

2. 2 Variants of pragmatemes

In the cases when the signifier of a pragmateme varies (on a syntactic, lexical, morphological or phonetic level) we speak about its *variants*, for instance:

Fr. *Ne pas donner de la nourriture aux animaux* (Do not give food to the animals) or *Ne pas nourrir les animaux* (Do not feed the animals) [on a sign in the zoo]. Here we have a periphrasis *nourrir* (to feed) – *donner de la nourriture* (to give food).

Rus. When someone sneezes, in Russian there are two manners to react according to the relationship with this person: *Будь здоров!* (Be healthy) is informal and *Будьте здоровы!* is formal (Be healthy) [to someone sneezing]. The difference is morphological – a formal form needs appropriate suffixes.

However, these expressions will always have a high level of restriction. For instance, when someone goes travelling, in French people say *Bon voyage!* (Good trip). Such grammatically correct phrases as #Joyeuse route (Happy way) or #Joyeux voyage (Happy trip) would be considered unacceptable by French native speakers. It exemplifies the phraseological character of pragmatemes.

3 Contrastive Study of Pragmatemes

During our research we collected an empirical data of 1000 pragmatemes in Russian, French and Spanish. It gave the possibility to compare equivalents in three languages, their semantic representations and the peculiarities of their usage. We revealed several differences in equivalent pragmatemes, detected pragmatemes specific for some languages and thus lacunas in others.

3. 1 Semantic differences

For the same situation and conceptual representation we encounter pragmatemes with different semantic representations. For instance, in France a sign near a lawn the expression prohibiting the walk on a lawn says *Respecter la pelouse* (Respect the lawn) whereas in Spanish it is *Prohibido pisar el césped* (It is prohibited to walk on the lawn). The Russian equivalent is *По газонам не ходить* (On the lawn not to walk). The difference consists of the choice of verbs to express the prohibition of walking on a lawn: in French – respecter (to respect), in Spanish – prohibido (is prohibited) and in Russian – не ходить (not to walk).

The Russian pragmateme which denotes that a bus does not take passengers and goes to the depot is *Автобус идет в парк* (The bus is going to the depot). In France and Spain the corresponding text is Fr. *Hors service*, Sp. *Fuera de servicio* (Out of service).

These are examples of the interlingual synonymy.

3. 2 Pragmatic differences

Another difference concerns the usage of pragmatemes. In both French and Spanish there are greetings which mean "Good day". Although the French pragmateme *Bonjour!* is used from 5 a.m. to 5 p.m., the Spanish equivalent *¡Buenos dias!* (Good day) is used until midday and then *¡Buenas tardes!* (Good afternoon) is used. In Russian the pragmateme *Добрый день!* which literally corresponds to "Good day"or "Good afternoon" is used from 11 a.m. to 5 p.m. In the morning (from 5 a.m. to 11 a.m.) the greeting is Доброе утро! (Good morning) is used. Therefore the French *Bonjour!* has the largest temporal usage among these three languages.The description of these of these expressions needs additional pragmatic information, like time of the use.

3. 3. Specific pragmatemes

The contrastive study of pragmatemes reveals the existence of expressions specific to some languages and of course cultures. For example, the French pragmateme found in trains *En cas d'affluence ne pas utiliser les strapontins* (In the case of rush hours do not use strapontins)[on a sign, in a carriage]. There are no strapontins neither in Russian nor in Spanish trains and that is why it is a lacuna in both languages.

Rainy weather and frequent mud in the streets is the source of the Russian pragmateme *Пожалуйста! Вытирайте ноги* (Please! Wipe your feet). This sign appears on the doors of shops, banks, schools, universities during cold, rainy or snowy seasons. We did not find an equivalent neither in France nor in Spain.

The Russian greeting to a person who has just taken a bath *С легким паром!* means approximately "I hope you have had a nice bath". This expression is specific to Russian culture and refers to the old tradition of taking a steam bath (banya). This pragmateme has no equivalent in either French or Spanish.

There are no symmetric languages with perfectly corresponding equivalents. The examined examples show that sometimes, in the same situation, different languages use expressions with different semantics. Sometimes there are some pragmatic differences in the use of pragmatemes. Specific cultural pragmatemes present great interest for linguists in contrastive studies. The provision of all peculiarities of pragmatemes is very important especially in the teaching and acquisition of foreign languages.

4 Pragmatemes Dictionary

As we have previously shown, pragmatemes have many semantic and cultural nuances. Their functioning is rather complex and depends on many pragmatic factors.

The knowledge of stereotypical utterances such as greetings, civilities, signs and posters, notices on packages is required by the initial levels of foreign language acquisition, in particular, levels A1 and A2 of language proficiency proposed by the Common European Framework of Reference for Languages (Differents authors, 2000).

Pragmatemes are also indispensable when you travel to another country and need to know what exact expression to use in a given situation. Thus, a non-native speaker risks making mistakes when formulating a correct answer to the French *Merci*. This is due to the existence of at least fourteen possibilities to answer according to the situation, relations between speakers, degree of politeness and etc.

4. 1 Works containing pragmatically bound expressions

We find useful expressions in phrasebooks and special linguistic sections in travel guides. Unfortunately, the majority of them does not give the necessary information for an adequate usage and correct comprehension of pragmatemes (Blanco, to appear). These books are not structured properly. Expressions are stored in the form of lists without any classification. Some of them do not even have a description of the main everyday situations. Linguistic data is often limited to an imitation of pronunciation. In order to make them a useful tool these issues need to be elaborated or improved by specialists in lexicography, translation or foreign language teaching.

As for a special pragmatic dictionary, the idea is not new. The Hungarian linguist Ivan Fonagy (1982) was the first to put forward this idea. He proposed to fill in the lacunas in traditional dictionaries by putting prefabricated expressions used in specific situations.

Michel Martins-Baltar developed this idea and created the first pragmatic monolingual dictionary of usual expressions bound by motives of utterance. The French dictionary DICOMOTUS (*Dictionnaire des expressions de motif usuelles*) is based on the model Motive <=> Reaction (Martins-Baltar, 2000). We find the information about the motive, the function, the conditions of the action presented by the expression, its form and its frequency. However some information seems to be superfluous, its cumbersome dimensions make it difficult to use especially for a non specialist.

With regards to bilingual dictionaries we can note The French-Romanian communication guide by Andrei Gancz (Gancz, 2006). Unlike traditional phrasebooks the Guide does not contain thematic vocabulary, but utterances related to speech acts and has a thematic structure. Its advantage is that it can be used for foreign language learning purposes.

Unfortunately, until now there has been no dictionary which contains pragmatemes and which has a detailed description of their situation of use.

4. 2 French-Spanish-Russian pragmatemes dictionary

During the two years (2009-2011) studying on the Erasmus Mundus International NLP & HLT Masters Programme we have developed a trilingual dictionary with pragmatemes in French, Spanish and Russian. Pragmatemes in the three languages do not refer to the translation of each concept but equivalents corresponding to the same situation.

The French-Spanish-Russian Pragmatemes Dictionary is based on the text production model developed in the scope of the Meaning-Text Theory. That means from a Concept(ual) R(epresentation) [= ConceptR] of the situation the speaker builds the Sem(antic) R(epresentation) [= SemR] of his future utterance. From the SemR he makes the utterance according to the language rules (Mel'čuk, 1998).

Pragmatemes stored in the Dictionary (1000 units in total) have the same structure of articles. The microstructure contains the information about the conceptual representation (a reflection of the reality that the speaker wants to express), the extralinguistic situation in which a pragmateme occurs (the corresponding pragmatic conditions of the use), its variants and synonyms. The pragmateme equivalents are given in three languages: French, Spanish and Russian.

Due to the lack of space we will give an example of French pragmatemes presented in the Dictionary, in group 9) Food and drink and its sub-group 9.1) Food packaging (See Table 1).

Conceptual representation	**Situation of communication**	**Pragmateme in French**	**Variant of pragmateme**	**Synonym of pragmateme**
Il faut conserver le produit au frais (It is necessary to keep the product cold)	[sur un emballage] (On the package)	*Conserver au frais* (Keep in cold)	*Garder au frais* (Save in cold)	*A conserver à l'abri de la chaleur* (Keep away from heat)

Table 1: Example of French pragmatemes in the French, Spanish, Russian Pragmatemes Dictionary

In traditional dictionaries entries are arranged in the alphabetical order. This approach is good for the reception: a learner consults these dictionaries when he does not know the meaning of a word. However, when a learner wants to find an equivalent for a pragmateme with the same conceptual representation in the same situation dictionaries with a semasiological order will not suit their needs.

Our dictionary has an onomasiological classification. That means that pragmatemes are classified according to the conversational topics: from a conceptual representation to a text of pragmateme. The Dictionary is conceived for foreign languages teaching/learning and includes the main topics studied on initial and intermediate levels of language acquisition.

All expressions are divided into 14 thematic groups:

1. Social interaction
2. House and environment
3. Holidays
4. Entertainment
5. Transport
6. Travelling
7. Health and well-being
8. Shopping
9. Food and drink
10. Places
11. Services
12. Correspondance
13. Education
14. Army

The proposed model of the dictionary is focused, firstly, on the foreign language learner. We hope that this dictionary would serve not only from time to time to explain the conceptual representation of the pragmateme but also as a useful tool for learning with the aim of

providing more autonomy to learners. Furthermore we expect that the dictionary of pragmatemes would also be useful for teachers. They would find ideas for organising and structuring the expressions during classes. Finally, the aim of the dictionary is to assure the development of production, reception and interaction skills of the student.

5 Conclusion

We have presented the phenomenon of pragmatemes and several pragmatic criteria that can be useful in its formal description. We have stated that a pragmateme can be represented by a single lexical unit, a phrase or a sentence. We have shown the interest of a contrastive approach and have revealed several semantic, pragmatic and cultural differences in Russian, French and Spanish pragmatemes. We have highlighted the importance of stereotypical utterances in foreign languages to learners and tourists travelling to a country where they do no speak the language. Neither phrasebooks nor pragmatic dictionaries give exhaustive information about pragmatemes, their meaning and their situation of utterance.

We have proposed a model of the Dictionary for the teaching/ learning of pragmatemes. It contains all pragmatic information for a correct comprehension and an adequate use of pragmatemes. The macrostructure of the dictionary focuses also on the learners' needs: 14 communicational topics help a non-native speaker to find out the right expression for the given conceptual representation. Variants and synonyms of pragmatemes complete the microstructure of the dictionary.

Acknowledgements

This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT Programme.

I would like to firstly acknowledge my principal research dissertation project supervisor Xavier Blanco Escoda (Universitat Autònoma de Barcelona) for the suggestion of the research topic, his useful advice and availability. Secondly, I am grateful to my co supervisor Izabella Thomas (Université de Franche-Conté) for her comments, support and kindness.

Bibliography

Blanco, X. to appear. Traduction des pragmatèmes dans les guides de conversation en russe. Contenus conceptuels et enjeux culturels. In Sfar, I. (ed.). *Synergies Tunisie,* Gerflint.

Differents authors. 2000. *Un cadre européen commun de référence pour les langues: apprendre, enseigner, évaluer*. Strasbourg: Conseil de l'Europe.

Fonagy, I. 1982. *Situation et signification*, Amsterdam: Benjamins.

Gancz, A. 2006. *Guide français-roumain de la communication*. Projet BABEL: http://projetbabel.org/roumain/guide.php

Martins-Baltar, M. 2000. Construire un dictionnaire d'énoncés (unilingue comme plurilingue): à propos de Dicomotus. In Szende, Th. (ed.). *Approches contrastives en lexicographie bilingue*, 301-317. Paris: Champion.

Mel'čuk, I., Clas, A., Polguere, A. 1995. *Introduction à la lexicologie explicative et combinatoire*, Louvaine-la-Neuve: Duculot.

Mel'čuk, I. 1998. Collocations and Lexical Functions. In Cowie, A. P. (ed.). *Phraseology: Theory, Analysis and Applications, (Oxford Studies in Lexicography and Lexicology)*, 79-100. Oxford University Press.

Mel'čuk, I. to appear. *Semantics: From Meaning to Text*, Amsterdam/ Philadelphia: Benjamins.

MTT meets Construction Grammar: the Treatment of Argument Structure

Daniel Weiss

Slavisches Seminar der Universität Zürich
Plattenstr. 43, CH-8032 Zürich
dawe@slav.uzh.ch

Abstract

Construction grammar offers a valuable tool for handling certain tricky cases that provide a challenge to MTT. In this paper, two such cases are examined, viz. 'hybrid' uses of non-motion verbs in motion constructions and omissions of motion verbs resulting in verbless sentences. Finally, the procedure of how to incorporate a construction into the MTT framework is substantiated by means of verbs denoting a physical transfer (giving).

Keywords

valency structure, Construction Grammar, verb omission, motion verbs, zero construction

1 Introductory remarks

When I first came across Construction Grammar, I could not detect anything impressing in it. After all, for somebody familiar with MTT, Ch. Fillmore's revelation that there exists a broad zone of semi-regular and non-compositional phenomena in between syntax and phraseology did not come as a surprise since Russian linguists had been studying such phenomena quite intensely long before the rise of CxG. Not only had part of the apparatus of Lexical functions been designed to cope with them, but terms such as 'syntactic phraseme' had been coined to the same purpose, and the strict division of labour between the grammar and the dictionary was already blurred by Apresjan's 'small syntax' or 'grammar of the dictionary'. Only on reading Goldberg's seminal work on argument structure (Goldberg 1995) I realized that CxG could perhaps offer an effective tool to capture certain generalisations about the syntax and semantics of huge verb classes that could not be handled in an appropriate way within the MTT framework. In particular, this holds for hybrid and often occasional argument structures resulting from the merger of two different constructions such as *He sneezed the napkin off the table, The truck rumbled down the street* or *I cannot imagine my way through the dark labyrinth of its distortion* (Goldberg 1995: 9-10). An MTT adherent will have a hard time

when analysing the structure of such sentences since there is no LF or Surface-syntactic relation that would do the job, nor would it make sense to add new meanings to the verbs involved in the dictionary. In this respect, CxG may be said to be more flexible than MTT in that it accounts for the syntactic elasticity of natural languages.[1] Therefore, it seems advisable to import at least some elements from CxG into MTT, without any risk of an "unfriendly take-over" of the latter by the former.

Before tackling our subject, some terminological remarks may not be out of place. To begin with, "a construction is posited in the grammar if it can be shown that its meaning and/or its form is not compositionally derived from other constructions existing in the language" (Goldberg 1995:4). Due to this all-embracing definition, virtually everything (from a morpheme up to an idiom or a clause) is an instance of a construction (Goldberg 2006:5). The type of construction to be analysed here can be best described as a generalized argument structure (roughly equivalent to a government pattern in MTT) plus some abstract meaning representing a class of verbs; for example, the central sense of the English ditransitive construction is defined as 'X CAUSES Y to RECEIVE Z', which is reminiscent of the left half of a semantic explication in an MTT dictionary. Moreover, appropriate constraints are added in order to filter out inacceptable combinations. Verbal lexemes may now be inserted into such a construction if they meet its argument structure and the constraints. Different clauses may be combined according to inheritance links between the corresponding constructions (see below, ex. 2, 6, and 8). Most crucial is the non-derivational character of all constructions: unlike MTT (but like many other contemporary syntactic theories), *John gave an apple to Mary* is not derived from *John gave Mary an apple*; this point cannot be elaborated here due to the lack of space. CxG arguments correspond to MTT Semantic actants in a given construction, whereas CxG participants may be equated with Semantic actants in the dictionary. MTT Circumstantials roughly coincide with adjuncts in CxG, and theta roles (semantic cases) such as agent or recipient are called roles in CG.

2 Three case studies

Goldberg (1995: 11) presents a series of different uses of the verb *kick*, all of which require different translations in Russian:

(1) Pat kicked the wall. *Pat pnul nogoj stenu / udaril nogoj po stene.*
(2) Pat kicked Bob black and blue. *Pat ispinal Boba nogami do sinjakov.*
(3) Pat kicked the football into the stadium. *Pat zakinul mjač na stadion.*
(4) Pat kicked at the football. *Pat udaril mjač nogoj.*
(5) Pat kicked his foot against the chair. *Pat pnul stul nogoj / udaril nogoj po stulu.*
(6) Pat kicked Bob the football. *Pat kinul / brosil / pasoval mjač Bobu.*
(7) The horse kicks. *Lošad' ljagaetsja.*
(8) Pat kicked his way out of the operating room. *Pat probilsja pinkami iz operacionnogo zala.*

This series serves ideally to illustrate the author's overall approach: "The verb is taken to be an *n*-place relation "waiting" for the exactly correct type and number of arguments". In the

[1] As is stressed in (Raxilina 2010:50-58), this does not hold for Apresjan's 1967 (and hence pre-MTT) monograph "Experimental'noe issledovanie russkogo glagola", where the author anticipated many of the fundamental ideas of CxG.

case at hand, she posits eight distinct argument structures. Moreover, examples 2, 6 and 8 provide instances of what she calls inheritance[2] links of constructions: the transitive *Pat kicked Bob black and blue* would be analysed as X DIRECTS ACTION at Y, and the whole clause inherits an additional argument from the resultative construction X CAUSES Y to BECOME Z. In 6, the construction X CAUSES Y to MOVE [to] Z inherits from the ditransitive construction X CAUSES Y to RECEIVE Z an additional argument for the recipient *Bob*. And finally, 8 presents an insertion of the verb X KICKS Y into the intransitive motion construction X MOVES [to] Y. For lack of space, details will not be discussed here. However, it should be emphasised that due to her focus on Construction syntax, (Goldberg 1995 and 2006) does not offer full-fledged lexical entries. In this way, we may but surmise what the lexical description of the 'rich semantics' of *kick* looks like; in particular, it remains unclear whether or to what degree polysemy has to be avoided in the dictionary.[3]

For an MTT adherent, these multifarious English uses of *kick* would be analysed in terms of polysemy, changing government patterns and Lexical functions. The details remain disputable; for instance, should the recipient *Bob* in 6 be assigned the role of an actant or a circumstantial (free adverbial, adjunct) as, e.g., in *Mary baked Jim a cake*? Is the resultative phrase *black and blue* in 2 to be treated as an instance of a separate Lexical function or just as a resultative subtype of the LF Magn?[4] And should we consider 8 a kind of syntactic phraseme with a frozen part (*his way* has to be coreferential with the subject) and two open slots for the indication of the agent and the source, path and/or the goal of the movement? As for 7, this use of *kick* is probably best handled as an instance of lexical conversion.

The Russian equivalents show a striking divergence with the English originals: instead of one single verb, we find not less than eight different lexemes (let alone their possible polysemy) and one paraphrase (example 8). The overall pattern is pretty obvious: the closest equivalent of *kick* in its primary meaning would be *pnut'*, which, however, turns out to be a mismatch in six out of eight sentences. Instead, informants prefer either the semantically less rich *udarit'* 'hit' and the synonymous pair *kinut'* / *brosit'* 'throw', or they employ prefixed verbs (*ispinat', zakinut'*) or a nominalisation (*pinkami*); finally, the kicking horse is rendered by a specialised verb (*ljagat'sja*). In other words: in neither case do we use a syntactic extension of the type described by Goldberg. This is most striking in example 6 which exhibits a pattern widespread in English (a caused-move construction is changed into a ditransitive construction): the litteral Russian equivalent **Pat pnul mjač Bobu* would be inacceptable, the appropriate verb being *kinut'* which allows for a dative in its government pattern both in literal and figurative meaning, cf. *Kin'te mne spasatel'nyj krug*! 'Throw me the life-saver!', *Ona kinula mne pis'mo* 'She sent me a letter'. As for the insertion of a verb of motion into a resultative construction illustrated by 2, Russian also offers a specialized word formation device: e.g. *Dan talked himself blue in the face* (Goldberg 1995: 9) would yield a translation with the

---

[2] For details, see Goldberg (1995 : 72-84). It goes without saying that this term has nothing in common with Mel'čuk's inheritance principle, see (Mel'čuk 2004 : 12).

[3] In Goldberg (2006 : 42) the author states that "*Kick*, for example, only has two profiled participant roles; the recipient argument in *She kicked him the ball* is added by the construction." This would be perfectly in line with my assumption that in the MTT approach *him* should be treated as a circumstantial (see below).

[4] The 'canonical' value of the LF Magn in the given case would be an intensifier such as *mercilessly*, or, in Russian, *bespoščadno*, an idiom like *bit' ne žalet'*, etc.

prefixed reflexive verb[5] *dogovorit'sja/doboltat'sja do*... Thus in all cases examined, Russian does not manifest the same syntactic elasticity as English simply because it does not need it: it may rely solely on lexical differentiations due to its richer verbal morphology (prefixation, reflexivisation). All this boils down to the statement that these examples can be easily handled in the Russian dictionary so that no need for a special construction grammar-like component arises.

Can these contrastive findings be generalized? English sentences such as *She sneezed the napkin off the table* quoted in the beginning likewise do not lend themselves to a Russian translation in one single sentence but have to be split up into two clauses. There exist, however, certain groups of Russian verbs that show the same behaviour as Goldberg's English examples. As is pointed out by (Raxilina 2000: 375), the following verbs denoting various sounds can combine with adverbials associated with the idea of motion:

(9) Diližans uxal / xljupal / skripel čerez derevnju 'the carriage groaned / gurgled / creaked through the village'

This is reminiscent of English examples such as *The truck rumbled down the street* discussed by Goldberg. Such cases should be strictly kept apart from verbs where both motion and sound production are equally part of the meaning proper, cf. *splash* as in *She splashed through the water*. In the MTT approach, one might be tempted to assign the PP *čerez derevnju* in (9) the status of a circumstantial, but against the implied sentence *Diližans exal čerez derevnju* this looks pretty counterintuitive. On the other hand, nobody would probably be inclined to posit a second meaning and a different government pattern for a verb like *skripet'* and treat it as a motion verb with an additional valency for the path; therefore, MTT has to shape a different solution for such cases. The same holds for verbs that describe the manner of movement by means of metaphorical physical activity (Raxilina 2000: 374):

(10) Diližans pilil / česal / molotil čerez derevnju 'the carriage cut (litt. sawed) / bolted (litt. scratched) / threshed through the village'

Again, there is no sense in creating separate dictionary entries for these uses unless we agree that they function as conventionalized metaphors and as such regularly govern spatial prepositions. Thus we may conclude that even in an MTT framework we do need some equivalent of a goal-directed motion construction that can account for such transitions of non-motion verbs into verbs describing both an intransitive motion and its accompanying circumstances such as the sounds produced by it or a characterization of its speed, etc. It goes without saying that we must equip this device with appropriate filters that will rule out, e.g., the following series of examples (all after (Raxilina 2000: 375)) **Poezd svistel čerez derevnju* 'The train whistled through the village', **Mal'čik govoril čerez derevnju* 'The boy spoke through the village'. As Raxilina points out, the verbs involved here denote communicatively meaningful sounds and therefore can no longer combine with mere verbs of motion. It could be added that in the case of *svistet'*, the information transfer arises only as a secondary,

[5] Note that the reflexive *himself* in the original wording is motivated otherwise: unlike in Russian, it does not function as a word formational device but signals the coreference of patient and agent in a transitive sentence, cf. English nonreflexive resultatives such as example 2 above.

derived meaning,[6] whereas in the primary meaning the concepts of sound and motion are perfectly compatible, cf. *Veter svistel čerez pustynnye polja* 'the wind whistled through the empty fields', *Par svistel čerez otverstie* 'the steam whistled though the vent': these examples are comparable to example 9 in that the whistle does not communicate anything,[7] moreover, the nouns 'wind' and 'steam' evoke themselves the idea of motion. Of course, the two concepts of motion and information transfer are perfectly compatible if the path is expressed explicitly by some circumstantial, cf. *Vsju dorogu čerez derevnju oni boltali, peli, peredraznivalis'*. Therefore, the real obstacle in the cases discussed above is the absence of any overt indication of a movement besides the preposition.

For the sake of completeness, let us add that the communicative event may itself be conceived of as a kind of abstract movement, cf.

(11) Komu on tam svistel *čerez plečo* Erika? 'Who did he whistle at over Eric's shoulder?' smotra.ru/clubs/1/blog/119256/

This question may refer to a situation where neither of the actors (Eric and the whistler) moved even one step; the only thing to be perceived as moving is the sound of the whistle. Again, the PP *čerez plečo Erika* has to be analysed as a circumstantial in an MTT framework.

Metaphorical uses of verbs denoting a physical action such as 10 require an atelic predicate (Raxilina 2000: 374), or, in Vendlerian terms, an activity: this accounts for the ungrammaticality of **Diližans vzmaxnul čerez derevnju* 'the carriage swept through the village' (*vzmaxnut'* denotes an achievement). Obviously, this is only part of the story: there must be a host of other restrictions at work which block for example verbs like *kovat'* 'forge', *šlifovat'* 'grind' etc. On the whole, it seems to be a fair assumption that this metaphorical type of derived motion verbs is much more restricted and more language-specific than the combination of sound and motion illustrated in 9 and should therefore be handled in the dictionary. Of course, this does not preclude a 'construction-like' component where the argument structure of intransitive motion verbs would be systematised.

To sum up, what we are looking for is a kind of generalized case frame (or else: government pattern) attributed to some abstract meaning of motion that would specify the conditions for the insertion of a non-motion verb. It seems to be advisable not to determine which subtypes of motion are involved in the given type of insertion. For example, the fact that sound-producing movements (cf. example 9) are rather associated with vehicles does not provide a sufficient motivation for creating a subtype that excludes walking.

3 Verb omissions

[6] This has a direct impact on its argument structure since now a valency for the addressee in the dative is added, cf. *svistet' sobake* 'to call the dog by a whistle'; this construction may also express disapproval, cf. *Publika mne svistela* 'the auditory booed me'. Note that such derived meanings otherwise tend to be realised by prefixes in Russian, cf. *osvistyvat'* 'boo'.

[7] Interestingly, German seems to be more flexible than Russian in this respect: cf. *Durchs ganze Dorf hindurch hat der Junge bloß geplappert* litt. 'the boy chatted through the whole village', which sounds acceptable to some speakers. However, if the bipartite adposition *durch…hindurch* is replaced by *durch*, the example becomes meaningless.

There is also an independent reason why our abstract motion construction should be shaped in as general terms as possible. Many languages allow for the omission of motion verbs in certain situations. In German, Polish and Czech for instance, this occurs in the narrative register if the narrator wants to depict a chain of events in more lively colours, cf.[8]

(12) Za każdym razem efekt jest ten sam: *kot ø na drzewo, lub przez wąską szparę do piwnicy,* a ja zziajany … wracam na miejsce startu 'Every time the effect is the same: the cat [climbs] up the tree, or [slips] through a narrow crack into the cave, and I get back to the starting point, completely exhausted'
zapiskibronislawa.bloog.pl/kat,0,m,2,r,2010,index.html 23.8.2010

(13) Rankiem budził Siostrę i mnie, celując w nas gorącymi, pachnącymi bułeczkami... *My ø do szkoły, On ø do łóżka* 'In the morning he used to wake up my sister and me and aim at us with hot fragrant buns … We [went] to school, he [went] to bed.
www.idn.org.pl/fson/kwart18/strpl.htm 23.8.2010

As can be seen, the variable in question must be some abstract pro-verb since in 12 we would need two different overt verbs to fill in the gap. The velocity of the sequence varies: in 12 we are rather dealing with a rapid course of action, in 13 the speed remains unspecified. Moreover, both examples demonstrate that this technique does not necessarily refer to single events but may also involve habitual activities. And finally, the sequence often involves two or more actors whose interplay may be best characterized in terms of action and reaction. In the discourse register, the omission of motion verbs is restricted to set phrases such as *A ty do kogo ø?* 'Who do you want to go to?' Contrary to this, colloquial Russian freely allows for verb omission in the discourse register, cf. *Ty kuda ø bez šapki*? 'Where [are you going / running / dashing etc.] without your cap?', *Ja ø v teatr zavtra* 'I [will go] to the theatre tomorrow', *Ja vot tol'ko čto s ulicy* 'I just [came in] from the street', *Xorošo by ø pod duš* litt. 'It would be great [to go] under a shower'.[9] The exact meaning of the missing verb (e.g. the manner of motion) is most often left open, and the same holds, as is shown by the examples, for grammatical categories (tense and mood). Narrative uses are of course also attested, cf. *Potom ø$_1$ drugoj kostjum i bystro ø$_2$ na scenu* 'Then he [changed into] another dress and quickly [returned] on stage' (Mažara 2010:236].

In the Russian tradition, such omissions are usually treated under the label 'zero motion verb'. This term is highly misleading since it suggests that what we are dealing with is a lexical unit; however, nobody has ever attempted to formulate an appropriate explication so far, let alone a full-fledged dictionary entry. In Weiss 2011a I have argued against the concept of zero verbs. I will not summarize the whole discussion here, but it should be emphasised that we would end up not with one, but at least three distinct 'zero motion verbs', since besides the type just examined, we find also contexts referring to undirected motion and (more importantly) to causation of motion. The former type may be illustrated by *Neudačnaja zima // Vot i na lyžax ø malo kak-to*[10] 'An unpleasant winter / I also did little skiing', the latter by *Ja takie pis'ma ø*

[8] For more Polish examples see Weiss (2011b, 2011c), for Czech parallels Mažara (2010) and MacShane (2000).

[9] Abundant material (though without sufficient context) is quoted in Širjaev (1973 : 299-304).

[10] Here, the missing motion verb would be spelled out as *katalsja* whose basic meaning is 'drove around'.

zakaznym vsegda 'I always [send] such letters by registered mail' (both examples from (Širjaev 1973)). What is more, we would have to cope with serious problems when delimitating the borders of this lexical hypercategory. For instance, how should we treat a missing verb such as *postupit'* in *Ona ø v prošlom godu v institut* 'She enrolled at the university last year'? Is this still a (metaphorical) motion verb, given that the primary meaning of *postupit'* is 'to step'?

The problems are aggravated by the fact that missing motion verbs are not the only group captured by the term 'zero verb' in Russian tradition. Other zeroes are postulated for verbs of communication, verbs of hitting/physical damage and physical transfer (giving). Hence, a considerable amount of overlap zones arises. To name but one: a verb such as *send* combines in one of its meanings 'caused motion' and 'physical transfer (cause to receive)' – but then the question arises which zero is represented in the abovementioned example *Ja ej takie podarki ø zakaznym vsegda,* a 'zero give' or a 'zero cause to move'? Note that this overlap is not a case of semantic ambiguity or vagueness, but merely of semantic cumulation. In a dictionary we must however posit distinct units without any overlap. We can of course create an additional, hybrid zero verb, named e.g. 'physical transfer by caused motion zero verbs'. This is the way chosen by (Wiemer 1996) who created a new hybrid group called verbs of addressing, which would account for the combination of motion and communication typical for these verbs. But when proceeding like this, we risk to end up with a proliferation of zero verbs that would still have more or less fuzzy boundaries. To mention a last candidate: (Širjaev 1973) proposed zero verbs called 'glagoly rečemyslitel'nogo dejstvija' (verbs of speaking and thinking') to cover such ambiguous cases as *Ja ø o drugom sovsem* 'I am speaking / thinking of something else'. One wonders what is won by creating such semantic monsters.

It is about high time to search for an alternative solution. What has Construction Grammar to offer in this respect? Recall what has been said above when discussing verbs of sound production employed as motion verbs: what we need is a generalized case frame (or else, argument structure) and some abstract complex of semes representing the meaning of motion. Now, all specialists of verb omission unanimously stress the importance of the remaining structure for the reconstruction of the missing verbal meaning. This remainder has to be composed by at least two elements; to be more precise, it should comprize at least one actant plus another actant or circumstantial (Weiss 2011a:142). A richer syntactic structure is realised in *Ja takie pis'ma ø zakaznym vsegda,* where four different constituents are involved. In many cases (including those quoted above), this syntactic remainder provides a case frame that allows to approximately identify the missing information. This is highly reminiscent of A. Goldberg's approach to Constructions grammar. Not surprisingly, the omission of motion verbs (but only these!) in Russian is mentioned as a separate construction in Goldberg (2006: 8) who quotes an unpublished manuscript by Chindarabam; her examples are *Kirill v magazin* 'Kirill goes/will go to the store' and *Kirill iz magazina* 'Kirill just got back from the store'.[11] But the Construction approach was already proposed before as a possible solution in Saj (2002: 107, 110, 111), who, however, did not elaborate his point. In addition, this author quotes Knjazev (2001: 35), who had anticipated the idea that in the case of zero verbs, not the

[11] This type is erroneously paralleled with the omission of the copula. Unlike missing motion verbs, the latter is restricted to the present tense, and there are a host of other arguments against the parallel.

verb requires the presence of certain case forms, but contrariwise "the presence of the latter evokes the vague image of a verb of motion, speech, physical action, etc." What has not been discussed so far in the pertinent Russian literature is Goldberg's concept of inheritance links (see above) which enables us for instance to account for such cases as 6 (*Pat kicked Bob the football*) in terms of two constructions, viz. CAUSE-MOVE and CAUSE-RECEIVE. In Russian, we may zero out the verb in a football reportage, obtaining for example : *Pat ø$_1$ Bobu, a tot ø$_2$ v vorota* ! 'Pat [passes the ball] to Bob, and Bob [hits] the goal'.[12] In this case, the missing information represents a combination of the abovementioned two constructions. The same holds for the omission of *poslat'* 'send' as in *Ja jej ø podarok včera* 'I sent her the present yesterday'. In a similar vein, the following ambiguity can be systematically predicted when traced back to Goldberg's analysis of the polysemy of the ditransitive construction:

(14) …no nam nužno ponjat', čto my polučim vzamen … Vot my xotim uznat' — a oni nam čto ø? '…but we have to understand what we'll get for it … We want to know what they [are offering/will give] us. (V. Putin)

Since the given argument structure fits into both constructions X CAUSES Y to RECEIVE Z (cf. give) and X INTENDS to CAUSE Y to RECEIVE Z' (offer), the missing verb can be reconstructed in both ways.

At this point it should be added that the predictive power of Goldberg's approach exceeds MTT in one respect: for sentences such as *She baked me a cake* or *Joe painted Sally a picture* which also exhibit the pattern of the ditransitive construction X INTENDS to CAUSE Y to RECEIVE Z, she formulates semantic constraints in order to single out deviant examples such as *Crush me a mountain*!, *Rob me a bank!* (Goldberg 1995: 36, 124-129, 141-151) and limit metaphorical extensions as in *The music lent the party a festive air.* In the MTT framework the "free" dative in these examples (except the last one) most likely represents a circumstantial (Mel'čuk 2004:279), but to my knowledge no MTT adherent has so far attempted to capture similar constraints in order to filter out inacceptable circumstantials.

In view of the abovementioned arguments it seems to be a sound proposal to introduce a limited set of zero constructions into MTT.[13] The appropriate place where to do this is undoubtedly the zone dubbed 'small syntax' ("Malyj sintaksis") by Ju. Apresjan. As early as in (Apresjan 1986: 63), he states: "some standard rules concern limited groups of verbs with neatly distinguishable common characteristics, for example identical syntactic or pragmatic features, coinciding constructions, etc." As far as I see, this proposal has never been applied to whole classes of verbs with common government patterns; instead, the 'small syntax' component has become the target domain for many detailed studies of syntactic idioms mostly carried out by L. Iomdin. In Weiss (2011a:152) I argued for the establishment of 'hyper-entries' (to be distinguished from ordinary dictionary entries) in this zone to cover zero constructions for the most salient types of verb omissions, i.e. verbs of goal-directed

[12] Besides this, there is a special construction restricted to press coverage, mentioned in Saj (2002 : 108): *Mostovoj na Karpina* 'Mostovoj [passes the ball] to Karpin'. In an MTT Framework, this would be an idiom with an own dictionary entry.

[13] As is argued in Weiss 2011, the major part of all verb omissions in colloquial Russian is contextually induced and cannot be described by means of a zero construction.

motion, communication, hitting, and giving ("glagoly predostavlenija"). Thus, most of the former 'zero verbs' would be treated like this. In the said paper, however, only the zero construction for verbs of giving was substantiated. The corresponding government pattern looks as follows:

Semantic actants	X = agent	Y = recipient / beneficiary	Z = object
Syntactic actants	1. S_{nom}	2. S_{dat}	3. S_{acc}

The semantic explication (*tolkovanie*) was not spelled out explicitly, but it seems fairly obvious that it should roughly follow the formula 'X CAUSES Y to RECEIVE Z'. This zero variant is realised if the surrounding context contains no lexical specifications typical for overt verbs, such as *dat'* 'give', *podarit'* 'donate, make a gift', *odolžit'* 'lend, borrow' or *prisudit'* 'award'.[14] All this is in line with Goldberg's approach. Yet, one is tempted to go one step further and include the zero construction into the overall intransitive construction (CAUSE-RECEIVE) with its core meaning of a "successful transfer between a volitional agent and a willing recipient" (Goldberg 1995: 151). This would allow accounting for all commonalities shared by the verbs of giving on the level of the 'small syntax'. Moreover, it would enable us to cover cases with a recipient used as circumstantial (cf. *She baked me a cake*) and to formulate the pertinent constraints. This would however imply that we also incorporate all other subtypes of the ditransitive construction distinguished in (Goldberg 1995: 38), notably X INTENDS to CAUSE Y to RECEIVE Z (cf. *leave, grant*), X CAUSES Y not to RECEIVE Z (cf. *refuse, deny)* and a group described as "Verbs of giving with associated satisfaction conditions" such as *guarantee, promise*. Note that Russian equivalents do not always fit into the case frame presented above: for example, *otkazat'* 'refuse' governs a prepositional phrase for the object (cf. *Direktor otkazal ej v podderžke* 'the boss refused her his support', and the two verbs *udostoit'* and *nagradit'*, whose meanings are pretty close to the "regular" *prisudit'* 'award', take a recipient in the accusative and an object in the genitive and instrumental, respectively.

If we turn now to the remaining candidates for Russian zero constructions, we recognize other important examples discussed by Goldberg, notably X MOVES [to] Y, X CAUSES Y to MOVE [to] Z and X TAKES ACTION at Y; to these should be added X COMMUNICATES with Y and Y COMMUNICATES with Y on Z, both of which play no significant role in her two monographs. If we proceed in the same way as with the verbs of giving, all abovementioned zero constructions will be considered instances of these constructions. The arguments in favour of this solution are the same as in the case of the verbs of giving: zero constructions exhibit the same argument structure as "normal" constructions with slots for overt verbs, and the semantics of the whole construction represents an abstract invariant that cannot be replaced by any existing overt verb. Moreover, the introduction of said constructions into the 'small syntax' would give us access to many fruitful generalizations, to mention but those concerning the transformation of verbs denoting sounds or physical actions into motion verbs illustrated in examples 9 and 10. Note that not only polysemy, but also homophony is not excluded in Construction Grammar; therefore, the fact that for example the case frame S_{nom} –

[14] This list of verbs raises the question of whether the meaning of the whole construction should not be equated with the basic meaning of the hypernym 'give'. This would however require a separate study.

do S_{gen} in Polish covers both X MOVES [to] Y and X COMMUNICATES with Y is not an obstacle.[15]

All this causes far-reaching consequences for the overall MTT architecture that cannot be discussed here. Obviously, the weight of the 'small syntax' component would significantly increase, and the role of the dictionary proper would have to be reexamined in order to limit redundancy. Polysemy remains crucial in both components: if Goldberg's constructions can be polysemous, does this reduce polysemy in the dictionary? As was mentioned in the beginning, Goldberg's work is of no help here since she carefully avoids formulating proper dictionary entries. But the decisive argument in favour of the incorporation of selected constructions into MTT is provided by a principle that has always been respected by MTT theoreticians: every generalization should be represented on an appropriate level of description, or else: if we agree that there are phenomena that cannot be adequately handled either in the dictionary or in the syntactical component, they have to be located in a separate module, i.e. the 'small syntax'.

Bibliography

Apresjan, Ju. D. 1986. Integral'noe opisanie jazyk i tolkovyj slovar', *Voprosy jazykoznanija* 2:57-69.

Goldberg, A. E. 1995. *A Construction Grammar Approach to Argument Structure*. Chicago & London: Routledge.

Goldberg, A. E. 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford: University Press.

Knjazev, Ju.P. 2001. Napravlenie ėvoljucii russkogo jazyka. In *Jazyki i obrazovanie. Sbornik naučnyx trudov,* 32-37, Velikij Novgorod.

MacShane M.J., 2000, Verbal Ellipsis in Russian, Polish and Czech, *The Slavic and East European Journal,* 44(2) :195–233.

Mažara Je. 2010. Swiss Cheese and the Lazy Speaker: The Omission of Verbs in Russian and Czech. in: Grønn, A. & Marijanovich, I. (eds.), *Oslo Studies in Language*, 2(1): 231-242.

Mel'čuk, I. 2004. Actants in semantics and syntax. I: actants in semantics, *Linguistics* 42(1):1-66, II.: actants in syntax, *Linguistics* 42(2): 247-291.

Raxilina, E. 2000. *Kognitivnyj analiz predmetnyx imen: semantika i sočetaemost'*, Moskva: Russkie slovari.

Raxilina, E. (ed.) 2010. *Lingvistika konstrukcij*. Moskva: Azbukovnik.

[15] In sentences with an omitted verb, this homonymy is however avoided in colloquial Polish; only in internet communication is it gaining ground. For details see (Weiss 2011b, c).

Saj, A. 2002. *Elliptičeskie konstrukcii: struktura i funcionirovanie* (na materiale russkogo jazyka). Diplomnaja rabota (unpubl.), Sankt-Peterburg.

Širjaev, E.N. 1973. O nekotoryx pokazateljax nezameščennyx sintakičeskix pozicij v vyskazyvanijax razgovornoj reči. In Zemskaja, E.A. (otv.red.) *Russkaja razgovornaja reč*, 191-273. Moskva: Nauka

Weiss, D. 2011a. Bezglagol'nye konstrukcii russkoj razgovornoj reči: ix tipologija i status v lingvističeskom opisanii. In Boguslavskij I.M. e.a. (eds.), *Slovo i jazyk. Sbornik statej k vos'midesjatiletiju Ju.D.Apresjana,* 139-155, Moskva: Jazyki slavjanskix kul'tur.

Weiss, D. 2011b. Dekapitacja zdań w polszczyźnie współczesnej. O konstrukcjach bezczasownikowych i ich odpowiednikach rosyjskich, *Poradnik Językowy* 1/2011 (Festschrift for A. Bogusławski), 111-121.

Weiss, D. 2011c. Pralinko, ja do Ciebie żartem, a Ty poważnie do mnie czyli o pozornej rusyfikacji polskiej komunikacji internetowej. *Prace filologiczne,* LX:299-319

Wiemer, B. 1996. Klassifikacija nulevyx skazuemyx v russkom jazyke po ix leksičeskim i referencial'nym xarakteristikam, *Studia z filologii polskiej i słowiańskiej*, 33: 245-273

Transfer of Russian Actantial Syntactic Relations into German

Robert Zangenfeind

CIS / Institute of Slavic Philology, Ludwig-Maximilians-Universität
Oettingenstraße 67, 80538 München, Germany
R.Zangenfeind@lmu.de

Abstract

Similarities and certain difficulties which arise in the translation of some Russian syntactic constructions into German are discussed. This concerns especially the necessity of transferring several Russian surface syntactic relations into different German relations in a number of cases. A list of German actantial surface syntactic relations is proposed.

Keywords

surface syntactic relation, dependency, transfer, (machine) translation, German, Russian.

1 Introduction

An essential feature of the linguistic processor ÈTAP-3, which was developed by Ju.D. Apresjan, I.M. Boguslavskij, L.L. Iomdin and others, is the machine translation from Russian to English and vice versa (cf. Апресян & al., 1989 and Apresjan & al., 2003). The machine translation from Russian to German is an additional feature which has a prototypical character until now. The level that is used for transfer between languages is the so called normalized syntactic representation.[1] The labels for surface syntactic relations which are used to describe dependencies in German in that prototypical version of ÈTAP are borrowed from English[2]. My aim is to introduce a system of syntactic relations for German which should be as similar as possible to the system of Russian relations in ÈTAP to have a basis for a computational implementation of Russian-German translation. Furthermore, I want to introduce German

[1] This is a level that is located between surface and deep syntactic representations. It uses, on the one hand, surface syntactic relations (which are specific for each language) to describe dependencies of words, but on the other hand e.g. Lexical Functions (which are independent of languages) as typical elements of the deep syntactic representation in MTT are used, and certain language specific words (e.g. auxiliary verbs, strongly governed prepositions) are removed at this level.

[2] Cf. the rules described in RA-TRADUCT.49 of ÈTAP rule manager in version 3.1.91 from the year 2008. This is part of ÈTAP-3 system which I am very grateful that Leonid L. Iomdin placed at my disposal.

labels for the German relations in order to take into account that the exact definitions of syntactic relations are specific for each language. In this paper, I will only consider German surface syntactic relations (= SyntRel) in the field of actantial relations.[3] In this field no new relations for German are needed—on the contrary, two of the Russian relations seem to have no counterpart in German (cf. sections 2.6 and 2.7)—whereas in the fields of attributive and auxiliary constructions some new relations will be necessary. In order to find out in which details German relations are different from Russian relations I found it useful to compare the syntactic constructions of Russian phrases and their German translations. As a result of this I will propose a list of German actantial relations[4] and consider some difficulties in transferring Russian syntactic relations into German relations. In doing so I will not present any ready-made rules, but simply show some possibilities of translations. In a lot of cases there are no (essential) changes necessary, so I will concentrate on some cases where changes are needed.

2 Transfer with some difficulties

(Апресян & al., 2010:24—31) describes seventeen Russian actantial relations. Seven of them, which entail some difficulties at least in some cases when transferred into German, are considered here in section 2. In what follows I'm not always proposing the best possible German translations of Russian phrases but translations that meet two requirements: they should have a syntactic construction as similar to the Russian original as possible and they should be at least acceptable or good German. These requirements are oriented at the needs of a possible implementation in machine translation.

2.1 Predicative Relation

In a lot of cases the Russian *предикативное синтаксическое отношение (предик)* 'predicative syntactic relation', which connects the predicate with its grammatical subject, is more or less equivalent to its German counterpart, the *prädikative syntaktische Relation (prädik)*.[5] Nevertheless, in some cases there are changes to be made: When there is no copula in Russian present tense the copula has to be inserted in German. This is rather trivial, nevertheless it has to be considered when constructing a dependency tree: there is the German *prädikative Relation* from the copula to the grammatical subject whereas there is the Russian

[3] Further relations, i.e. in the fields of attributive, coordinate and auxiliary constructions will be considered in future papers. I also will not deal here with the problem of word order in German; for a discussion on that, see e.g. (Gerdes & Kahane, 2007)

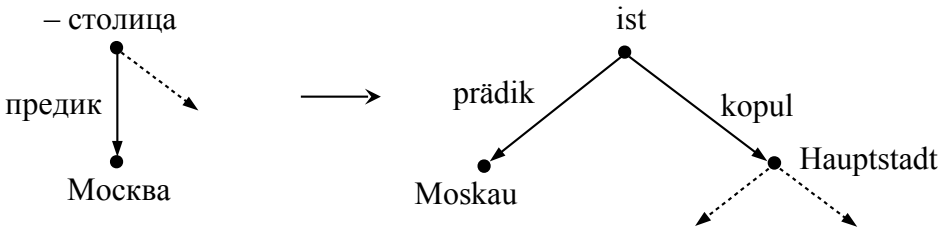
[4] For the names and definitions of German relations I follow in a first step the Russian and English relations that are used in ÈTAP, cf. (Апресян & al., 2010:24—43) and (Апресян & al., 1989:71—121), and then adapt them to the requirements of German syntax. German linguists, like e.g. (Heringer, 1996), often do not use any labels for syntactic relations, whereas (Specht, 2003) uses partially different labels. I, nevertheless, will use a terminology as close to ÈTAP as possible in order to facilitate a computational implementation.

[5] The definitions of these two relations, however, are different in detail: in the simplest construction, the Russian predicative relation connects a finite verb with the grammatical subject in the nominative case; this also holds for the German equivalent. In Russian constructions, however, the subject can also be e.g. a noun in the genitive or partitive case, which is not possible in German; cf. the definition of the Russian predicative relation in (Апресян & al., 2010:25). Differences of this kind are also to be found in the definitions of the other Russian relations and their German equivalents.

предикативное отношение from the nominal or attributive predicate to the grammatical subject.[6] The German equivalent of the Russian predicate is a dependent of the *kopulative Relation (kopul)* 'copulative relation', which connects the copula with its object (cf. section 2.4), as illustrated in (1a) with the dependency tree of (1b)[7]:

(1a) Москва – столица России.[8] '*lit*. Moscow – capital of Russia' → Moskau ist die Hauptstadt Russlands. 'Moscow is the capital of Russia'

(1b)



The broken lines represent relations to lexemes which are not relevant for the transfer of the Russian *предикативное отношение*.

In translating Russian verbs with the reflexive particle *-ся/-сь* which express passive voice, an auxiliary verb has to be inserted in German. This auxiliary is the syntactic head of the phrase; the participle of the full verb is connected to the auxiliary via the *passiv-analytische Relation (pass-anal)* 'passive analytical relation':[9]

(2) Заявка [Y] изучается [X]. '*lit*. The request audits-*reflexive*' – Der Antrag [Y] wird [X] [–pass-anal→] geprüft. 'The request is audited' [10]

Russian constructions with *есть* 'to be' as an existential verb have to be translated into a different German construction, e.g. with *es gibt* '*lit.* it gives, *i.e.* there is'; the expletive *es* 'it' is then grammatical subject (and thus the dependent of the *prädikative* Relation), the German equivalent of the Russian subject is dependent on the verb via the *1. kompletive Relation (1.kompl)* '1. completive relation', which connects a predicate with its first object:[11]

(3a) Есть внешний повод для этого. '*lit.* Is an external cause for it' – Es gibt einen äußeren Anlass dafür. 'There is an external cause for it'

---

[6] Cf. rule RA-EXPANS.08 of ÈTAP.

[7] In the dependency trees I will use inflected forms of lexemes for a better reading instead of the names of lexemes plus morphological characteristics that are standard in MTT.

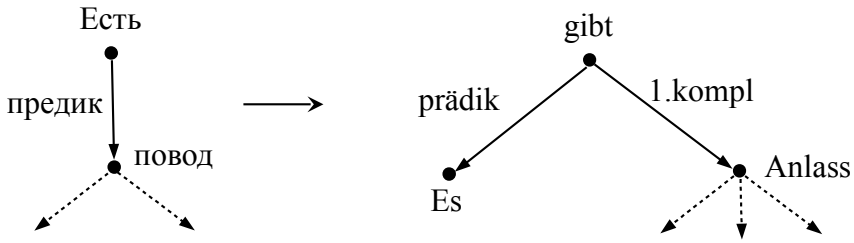
[8] Russian examples are taken from (Апресян & al., 2010) and from http://www.ruscorpora.ru.

[9] The *passiv-analytische Relation* is not an actantial but an auxiliary syntactic relation which is not treated here; cf. rule RA-EXPANS.54 of ÈTAP.

[10] Because of lack of space it is not possible to contribute dependency trees for all examples. To show the syntactic relation in question Latin letters are used to indicate the syntactic governor [X] and dependent [Y].

[11] Here we have a mismatch at the deep syntactic level between the two constructions, as they are described in (Mel'čuk & Wanner, 2006): the Russian construction includes a support verb $Func_0$, whereas the German construction is a rare case with the support verb $Oper_0$; cf. (Reuther, 2003).

(3b)

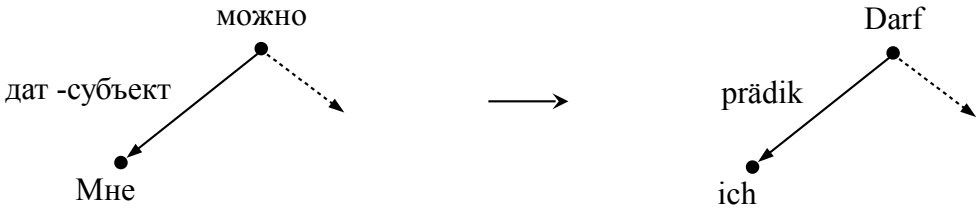


2.2 Dative-subject Relation

In certain cases no changes are necessary in the transfer of the Russian *дательно-субъектное отношение (дат-субъект)* 'dative-subject relation' which has the German counterpart *Dativ-Subjekt-SyntRel (dat-subjekt)*. Some Russian constructions, however, e.g. with *можно* '*lit*. it is possible, *i.e.* one can, one may' or *должно* '*lit*. it is obligatory, *i.e.* one should, one ought' or an impersonal modal verb entail a transfer of the *дат-субъект* relation into a German *prädikative Relation*. Russian *можно* (cf. example 4) or *должно* or the impersonal modal verb (cf. example 5) respectively is translated into a personal verb in German. The Russian dative subject is translated into a German nominative subject which is a syntactical dependent of the personal verb via the *prädikative Relation*:

(4a) Мне можно уйти? '*lit.* To-me it-is-possible to go?' – Darf ich gehen? 'May I go?'

(4b)

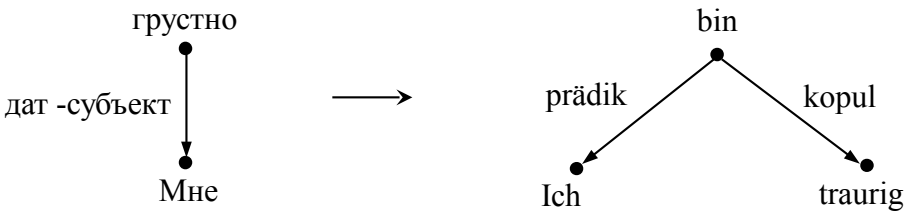


(5) Мне [Y via дат-субъект] пришлось [X] сказаться больным. 'To-me had-*reflexive* to report sick' – Ich [Y via prädik] musste [X] mich krank melden. 'I had to report sick'

In some cases when the Russian predicate is an adjective, the copula has to be inserted in German and with that a *kopulative Relation* 'copulative relation' to the adjective (cf. also (1)):

(6a) Мне грустно. '*lit*. To-me sad' – Ich bin traurig. 'I am sad'

(6b)



In some cases when in Russian there is no modal verb but an infinitive plus dative subject, a modal verb has to be inserted in German (e.g. *dürfen, können, müssen* 'may, can, must'). The full verb which is syntactic head of the subject in the Russian phrase depends on the modal verb in German via the *1. kompletive Relation*:

(7) Мне [Y via дат-субъект] сейчас выходить [X]. '*lit.* To-me go now' – Ich [Y via prädik] muss [X] nun gehen. 'I must go now'

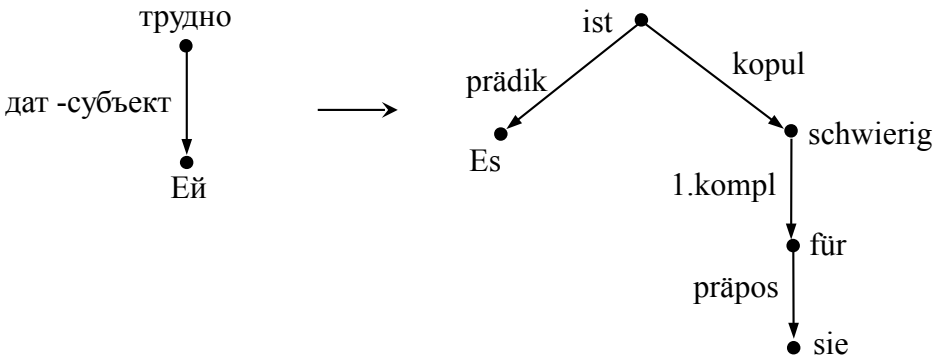
When translating some Russian constructions, the copula, and the preposition *für* 'for' and with that the *präpositionale Relation (präpos)* 'prepositional relation' (cf. section 2.5) have to be inserted in German.[12] This means that the Russian dative subject is translated into a German prepositional group where the preposition is a dependent of the adjective via the *1. kompletive Relation* and at the same time the head of the noun via the *präpositionale Relation*:

(8) Пете [Y via дат-субъект] это совсем неинтересно [X]. '*lit.* To-Petja this absolutely uninteresting' – Das ist völlig uninteressant [X] für [Y via 1.kompl] Petja. 'That is absolutely uninteresting for Petja'

Sometimes an additional expletive *es* 'it' as grammatical subject has to be inserted in German:

(9a) Ей трудно. '*lit.* To-her difficult' – Es ist schwierig für sie. 'It is difficult for her'

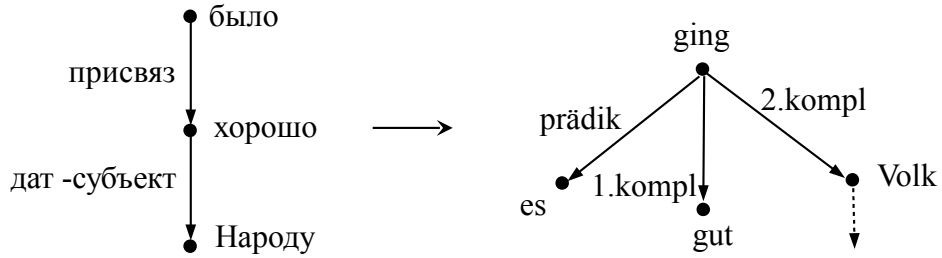
(9b)



In other cases a completely different syntactic construction is needed in German with *es geht* '*lit.* it goes' – a phrase which has two syntactic complements:

(10a) Народу было хорошо. '*lit.* To-people was well' – Dem Volk ging es gut. '*lit.* To-the people it went well; *i.e.* people felt well'

(10b)[13]



2.3 Agentive Relation

There are no big difficulties in the transfer of the Russian *агентивное СинтО (агент)* 'agentive relation' into the German *agentive Relation (agent)*. Some changes, nevertheless,

[12] To be precise, this preposition, being semantically empty, is not inserted at the transfer phase, but during the expansion from the normalized to the surface syntactic representation (cf. also rules RA-EXPANS.05 and RA-EXPANS.06 of ÈTAP). Nevertheless, since this paper is dealing with surface syntactic relations, I only will compare these levels of representation in both languages.

[13] The Russian *присвязочное СинтО (присвяз)* 'copulative relation' is the equivalent of the German *kopulative Relation* (cf. section 2.4).

are needed in all cases. If a verb is the head of the Russian *агентивное СинтО* 'agentive relation', in German there has to be inserted the preposition *von* 'from, by' or *durch* 'by'[14] which is then dependent on the verb via the *agentive Relation*. The preposition is the head of the noun via a *präpositionale Relation* (cf. section 2.5). The agentive Russian noun that is in the instrumental case will become a German noun in the dative case if the preposition *von* is inserted, or in the accusative case if the preposition *durch* is inserted (these details, of course, are defined in the government patterns of the two prepositions):

(11) рассматриваемый [X] комиссией [Y] вопрос '*lit.* considered by-commission issue' – die von [Y] der Kommission betrachtete [X] Frage 'the issue considered by the commission'

The Russian participle used in (11) alternatively can be translated into a German subordinate relative clause; but this does not have any further influence on the transfer of the agentive relation as described above – it only has influence on the translation of the participle itself:

(12) вопрос рассматриваемый [X] комиссией [Y] '*lit.* The issue considered by-commission' – die Frage, die von [Y] der Kommission betrachtet [X] wurde 'the issue that is considered by the commission'

In a sentence where a reflexive verb is used to express passive voice in Russian, an analytical verb form is used in German which again does not have any further influence on the transfer of the agentive relation; see the following example (cf. also example 2):

(13) Вопрос рассматривается [X] комиссией [Y]. '*lit.* The issue considers-*reflexive* by-commission' – Die Frage wird von [Y] der Kommission betrachtet [X]. 'The issue is considered by the commission'

If a noun is the head of the Russian *агентивное СинтО* 'agentive relation', this relation has to be transferred into a German *quasi-agentive Relation (quasi-agent)* because this better meets the definitions of these two relations.[15] In this case the government pattern of the noun defines which preposition (*von*, *durch* 'from, by' etc.) has to be inserted:

(14) приём [X] президентом [Y via агент] делегации '*lit*. the welcome by-president of the delegation' – der Empfang [X] der Delegation durch [Y via quasi-agent] den Präsidenten 'the welcome of the delegation by the president'

2.4 Copulative Relation

When a Russian copula is translated into a German copula, in a lot of cases there are no essential changes needed. When the dependent noun or adjective in Russian is in the instrumental

[14] Cf. footnote 12 and rule RA-EXPANS.07 of ÈTAP.

[15] I propose to use the German *agentive Relation* for verbs in passive voice and the *quasi-agentive Relation* for nouns, whereas in ÈTAP-3 the Russian *агентивное* 'agentive' relation is used both for verbs in passive voice and for nouns as syntactic heads if the dependent is a noun in the instrumental case; cf. the Russian definitions by (Апресян & al., 2010:26). Another possible approach would be to differentiate between nouns with a meaning of artefacts and others according to the definitions of the related English relations developed by (Апресян & al., 1989:77f.).

case, this dependent always will be in the nominative case in German. However, this does not have any influence on the syntactic relations and the Russian *присвязочное СинтО (присвяз)* 'copulative relation' is transferred into a German *kopulative SyntRel (kopul)*:

(15) Он был [X] учителем [Y,instr]. – Er war [X] Lehrer [Y,nom]. 'He was a teacher'

An important difference, however, between Russian and German is that there are more copulative verbs in Russian. In ÈTAP seven Russian verbs are marked with the syntactic feature 'СВЯЗ' as copula verbs, namely *бывать* 'to be, to visit, to happen, to take place', *быть* 'to be, to exist *etc.*', *делаться1* 'to become', *казаться* 'to seem', *оказываться* 'to turn out to be', *оставаться* 'to continue to be, to remain, to stay', *становиться1* 'to become'. In German there are only three copula verbs: *sein* 'to be', *werden* 'to become' and *bleiben* 'to continue to be'. Five of the Russian verbs can be translated into German copula verbs, whereas two of them, namely *казаться* 'to seem' and *оказываться* 'to turn out to be' are translated into German full verbs and thus the Russian *присвязочное СинтО* is transferred into a German *1. kompletive Relation*, cf. (16). This also holds for some translations of the five remaining Russian copula verbs into German verbs other than copula.

(16) Он казался [X] больным [Y via присвяз]. – Er wirkte [X] krank [Y via 1.kompl]. 'He seemed sick'

Depending on the verb that is used in German in some cases the adjective has to be extended to a construction with a full verb plus the particle *zu* 'to' and an infinitive of the copula; the predicative adjective then is a dependent of this newly inserted copula:

(17) Он казался [X] больным [Y via присвяз]. – Er schien [X] krank zu [Y via 1.kompl] sein. 'He seemed to be sick'

In some cases an expletive *es* 'it' is needed, which is a dependent of the copula via the *prädikative Relation*; this does not have any influence on the *kopulative Relation*:

(18) Жаль [Y] было [X] расставаться с ним. '*lit.* A pity was to part with him' – Es war [X] schade [Y], sich von ihm zu trennen. 'It was a pity to part with him'

2.5 Prepositional Relation

The Russian *предложное СинтО (предл)* 'prepositional relation' is transferred into its German counterpart, the *präpositionale Relation (präpos)* quite often. In some special cases, however, the Russian preposition is translated into a German adverb which entails a change of syntactic construction: instead of the Russian prepositional relation from the preposition to the noun, we find that the direction of the German relation, which is an *(eigentliche) attributive Relation (attrib)* 'attributive relation proper', is reversed, namely from the noun to the adverb:

(19) Спортсмен бегал по [X] 20 километров [Y via предл] в день. – Der Sportler lief jeweils [Y via attrib] 20 Kilometer [X] am Tag. 'The sportsman ran 20 kilometers a day'

In (20) we have a situation similar to (19): an appropriate translation again includes the transfer of the *предложное* 'prepositional' into a German *(eigentliche) attributive Relation* by

reversing the direction of dependency, but here a numeral is the syntactic head of the adverb *über* 'over'[16] via the *(eigentliche) attributive Relation*:

(20) Свыше [X] ста человек [Y via предл] явилось на субботник. – Über [Y via attrib] 100 [X] Menschen erschienen zum Subbotnik. 'Over a hundred people came to the subbotnik'

In (21) we again have an example where the Russian preposition is translated into a German adverb, but here the adverb is connected to the numeral as dependent of a *restriktive Relation (restr)* 'restrictive relation', which connects a word with a restrictive particle or adverb:

(21) около [X] десяти [Y via предл] – ungefähr [Y via restr] zehn [X] 'about ten'

2.6 Sentential-predicative Relation

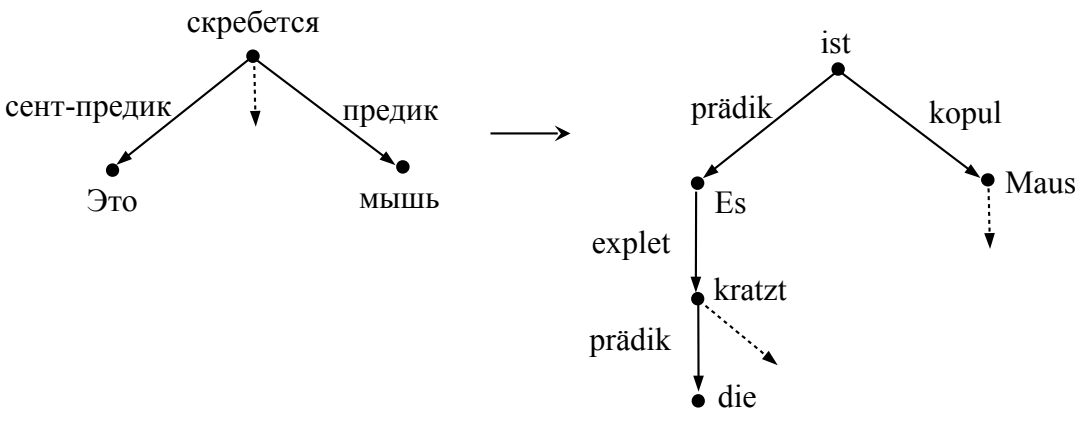
Changes in the transfer are always necessary when the Russian sentence contains a *сентенциально-предикативное (сент-предик)* 'sentential-predicative' relation because this relation seems to have no counterpart in German. In the simplest case the Russian lexeme that is the dependent of this relation (*это* 'it, that, this' or *то* 'it, that') is translated as German *hier* 'here' which is connected via the *(eigentliche) adverbiale Relation (adverb)* 'adverbial relation proper' as a dependent of the verb, cf. (22):

(22) Это [Y via сент-предик] разговаривают [X] друзья. – Hier [Y via adverb] sprechen [X] die Freunde. 'Here the friends are speaking'

However, in some cases the syntactic construction has to be changed radically. Russian *это* 'it, that, this' or *то* 'it, that' then is translated into a German construction that starts with *es ist* 'it is' if the Russian grammatical subject is a singular noun or *es sind* '*lit.* it are' if the Russian subject is a plural noun. The noun that is the grammatical subject in the Russian sentence is translated into a German noun that depends on the copula via the *kopulative Relation*. The expletive *es* 'it' is connected to the copula via the *prädikative Relation* and is the syntactic governor of the full verb in the relative clause via the *expletive Relation (explet)*:

(23a) Это мышь скребется за печкой. – Es ist eine Maus, die hinter dem Ofen kratzt. 'It's a mouse that scratches behind the stove'

(23b)



[16] The German *über* 'over' is used here as an adverb like e.g. *fast* 'almost' or *ungefähr* 'about'.

2.7 Presentative Relation

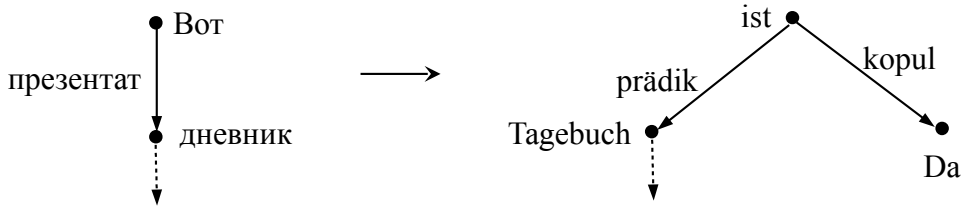
The Russian *презентативное СинтО (презентат)* 'presentative relation' also has to be changed always when transferred into German because again there seems to be no counterpart in German: Russian *вот* or *вон* 'there, here' respectively are translated into the adverb *da, hier* 'here' or *dort* 'there' and the syntactic construction is changed considerably. Instead of the Russian *презентативное СинтО* that has *вот* or *вон* 'there, here' as syntactic head and the verb as dependent, in German there is an *1. kompletive Relation* with inversed direction from the verb to *da, hier* 'here' or *dort* 'there':

(24) Вон [X] стоит [Y via презентат] дом. – Dort [Y via 1.kompl] steht [X] ein Haus.
'There stands a house'

When there is no verb in the Russian phrase, the copula verb has to be inserted in German. This copula in German is the head of the construction and the noun that is dependent on *вот* or *вон* 'there, here' in Russian, is dependent on the copula via the *prädikative Relation* in German. The adverb is connected to the copula via the *kopulative Relation*:

(25a) Вот мой дневник. – Da ist mein Tagebuch. 'Here is my diary'

(25b)



3 Transfer without essential changes

The ten remaining Russian actantial syntactic relations described by (Апресян & al., 2010:24—31) do not entail essential changes when transferred into German, as preliminary studies suggest. For these relations I would like to propose the following German counterparts:

- Квазиагентивное СинтО (квазиагент) → quasi-agentive SyntRel (quasi-agent)
- Несобственно-агентивное СинтО (несобст-агент) → uneigentliche agentive SyntRel (uneigent-agent)
- Первое комплетивное СинтО (1-компл) etc. → erste kompletive SyntRel (1. kompl) etc.
- Первое несобственно-комплетивное СинтО (1-несобст-компл) etc. → erste uneigentliche kompletive SyntRel (1. uneigent-kompl) etc.
- Неактантно-комплетивное СинтО (неакт-компл) → nichtaktantisch-kompletive SyntRel (nichtakt-kompl)
- Комплетивно-аппозитивное СинтО (компл-аппоз) → kompletiv-appositive SyntRel (kompl-appos)
- Подчинительно-союзное СинтО (подч-союзн) → subordinierend-konjunktionale SyntRel (subord-konj)
- Сравнительное СинтО (сравнит) → komparative SyntRel (kompar)
- Сравнительно-союзное СинтО (сравн-союзн) → komparativ-konjunktionale SyntRel (komp-konj)

- Элективное СинтО (электив) → elektive SyntRel (elektiv)

4 Conclusion

In the field of actantial syntactic relations ten Russian relations are supposed to require no essential changes which has to be verified in further studies. Seven relations require few modifications up to essential changes in constructions at least in some cases when transferred into German. This is especially the case when considering the two Russian relations which seem to have no counterpart in German, the *сентенциально-предикативное* 'sentential-predicative' and the *презентативное* 'presentative' relation but also holds for certain cases of the other relations considered. Of course, there are a lot of further problems in translating Russian actantial constructions into German which are not touched in this paper due to shortage of space. These more specific problems as well as attributive, coordinate and auxiliary syntactic relations are the subject of further research and will be considered in future papers.

Acknowledgements

Many thanks to Barbara Sonnenhauser and to three anonymous reviewers for their comments on the first version of this paper.

Bibliography

Apresjan, Ju.D. & al. 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning <=> Text Theory. In *Conference Proceedings of MTT 2003*, 279–288. Paris [http://proling.iitp.ru/publications].

Gerdes, K. & S. Kahane. 2007. Phrasing it differently. In Wanner, L. (ed.). *Selected Lexical and Grammatical Issues in the Meaning–Text Theory*, 297–335. Amsterdam, Philadelphia.

Heringer, H.J. 1996. *Deutsche Syntax. Dependentiell*. Tübingen.

Mel'čuk, I.A. & L. Wanner 2006. Syntactic Mismatches in Machine Translation. In *Machine Translation*, 20, 81–138.

Reuther, T. 2003. Support verb combinations with existential verbs (German and Russian). In *Conference Proceedings of MTT 2003*. Paris. [http://meaningtext.net/mtt2003/proceedings]

Specht, V. 2005. An MTT-Based Parser for German. In Ю.Д. Апресян (ed.). *Восток – Запад*. Москва, 316–329.

Апресян, Ю.Д. & al. 1989. *Лингвистическое обеспечение системи ЭТАП-2* . Москва.

Апресян, Ю.Д., И.М. Богуславский, Л.Л. Иомдин & В.З. Санников 2010. *Теоретические проблемы русского синтаксиса. Взаимодействие грамматики и словаря*. Москва.