Bilingual Dictionary Drafting: Bootstrapping WordNet and BabelNet

David Lindemann
UPV/EHU University of the Basque Country
david.lindemann@ehu.eus

Fritz Kliche
University of Hildesheim
fritz.kliche@uni-hildesheim.de

Overview

- Introduction
  - Motivation
  - Overview of Bilingual Dictionary Drafting (=BDD) methods
  - Some previous research
- BDD using concept-oriented lexical resources: The example of Basque-English
  - Concept-oriented vs. headword-oriented resources
  - Data extraction from WordNet / BabelNet: Workflow
- Basque-English dictionary draft: Evaluation
  - Standard Basque dictionary headwords
  - Quantitative Evaluation: BabelNet English-Basque intersection
  - Qualitative Evaluation: Assessment of translation equivalents
- Post-processing / editing issues
- Conclusions and Further Work

Motivation

- 400+ languages with 1 million L1-speakers or more
- Availability of bilingual dictionaries: Many scarcely resourced language pairs
  - Even where one of the top ten languages is involved
  - Example Basque: Only ES, FR, EN, RU, (DE) are covered
- Possible ad-hoc-workarounds for scarcely resourced language pairs:
  - (1) To use two bilingual dictionaries
  - (2) To use an automatically built dictionary or MT (more and more of them available)
  - Disadvantages
    - Time consuming
    - Mislead lookups (main problem: Polysemy / asymmetric lexicalization)
- Lexicography for uncovered language pairs (=from scratch)
  - Automated drafting of translation equivalent pairs
    - Saves human resources
    - Reciprocal bootstrapping: Upgrading of the resources employed for BDD

BDD Methods: A brief overview

- Corpus-based
  - Word alignment in parallel corpora
    - Bilingual parallel corpus: Bilingual word lists (without Word Sense Disambiguation)
      - Gale & Church 1991, Heja 2010, among others
    - Multilingual parallel corpus: Information for WSD using asymmetries in lexicalization across languages
      - cf. Lefever 2012, 2014 among others; see Kazakov & Shahid 2013 for a survey

BDD Methods: A brief overview

- Corpus-based
  - Word alignment in parallel corpora
    - Bilingual parallel corpus: Bilingual word lists (without Word Sense Disambiguation)
      - Gale & Church 1991, Heja 2010, among others
    - Multilingual parallel corpus: Information for WSD using asymmetries in lexicalization across languages
      - cf. Lefever 2012, 2014 among others; see Kazakov & Shahid 2013 for a survey
- Dictionary Pivoting
  - Connecting lemma-based lexical resources to each other
    - Filtering of polysemy related errors with corpus-based methods
      - cf. Saralegi et al. 2012, among others

BDD Methods: A brief overview

- Corpus-based
  - Word alignment in parallel corpora
    - Bilingual parallel corpus: Bilingual word lists (without Word Sense Disambiguation)
      - Gale & Church 1991, Heja 2010, among others
    - Multilingual parallel corpus: Information for WSD using asymmetries in lexicalization across languages
      - cf. Lefever 2012, 2014 among others; see Kazakov & Shahid 2013 for a survey
- Dictionary Pivoting
  - Connecting lemma-based lexical resources to each other
    - Filtering of polysemy related errors with corpus-based methods
      - cf. Saralegi et al. 2012, among others
- Bootstrapping concept-oriented resources
  - Wikipedia Interlanguage Links (cf. Navigli & Ponzetto 2010)
  - Open Multilingual WordNet (Bond & Foster 2013, cf. Varga et al. 2009)
  - ConceptNet (Speer & Havasi 2012)
  - BabelNet (Navigli & Ponzetto 2010)

Own previous research on BDD

- Lindemann et al. 2014 (Euralex Bolzano)
  - Set of (semi)-automatic methods for German-Basque bilingual word list building
  - Without Word Sense Disambiguation
  - Showcase German-Basque, an scarcely resourced language pair
    - Data for 2/3 of German 40,000 frequency lemma list, half of it accurate

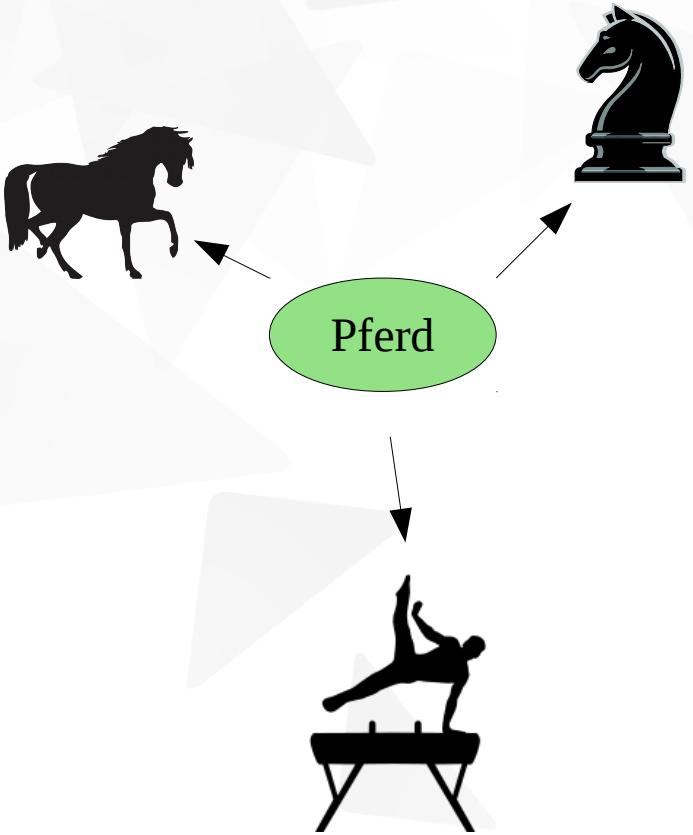
Own previous research on BDD

- Lindemann et al. 2014 (Euralex Bolzano)
  - Set of (semi)-automatic methods for German-Basque bilingual word list building
  - Without Word Sense Disambiguation
  - Showcase German-Basque, an scarcely resourced language pair
    - Data for 2/3 of German 40,000 frequency lemma list, half of it accurate
- Lindemann & San Vicente 2016 (Euralex Tbilisi)
  - Proposal of a lexicographic workflow for bilingual dictionaries with Basque
  - BDD including discrimination of homographous lemmata and word senses
    - Drafting of lemma list and lemma-POS-entities by bootstrapping Basque NLP resources
    - Linking to translation equivalents at word sense level via Princeton WordNet
  - Automatic and manual gap detection
  - Manually edited lexical data eventually sent back to Basque WordNet and other data providers

Own previous research on BDD

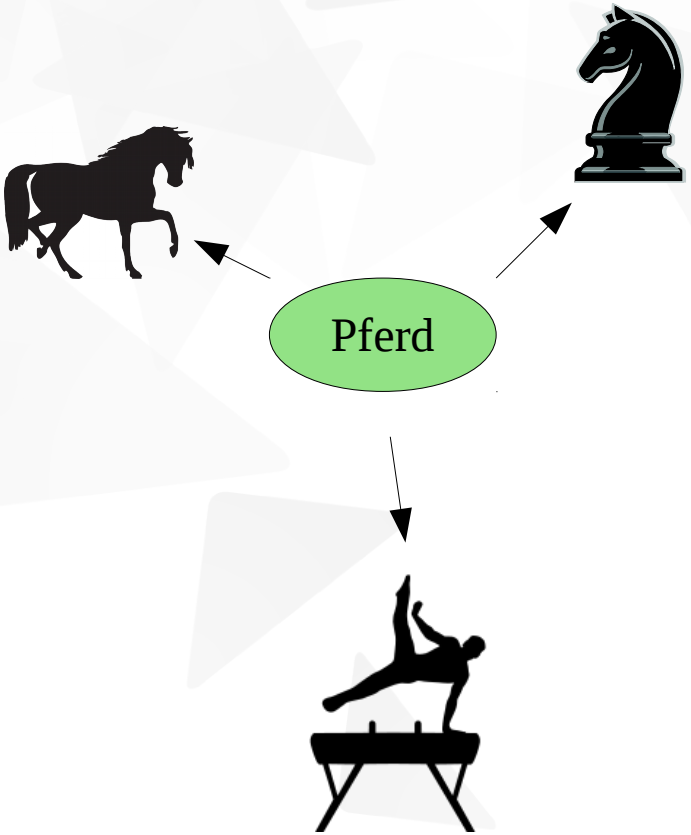
- Lindemann et al. 2014 (Euralex Bolzano)
  - Set of (semi)-automatic methods for German-Basque bilingual word list building
  - Without Word Sense Disambiguation
  - Showcase German-Basque, an scarcely resourced language pair
    - Data for 2/3 of German 40,000 frequency lemma list, half of it accurate
- Lindemann & San Vicente 2016 (Euralex Tbilisi)
  - Proposal of a lexicographic workflow for bilingual dictionaries with Basque
  - BDD including discrimination of homographous lemmata and word senses
    - Drafting of lemma list and lemma-POS-entities by bootstrapping Basque NLP resources
    - Linking to translation equivalents at word sense level via Princeton WordNet
  - Automatic and manual gap detection
  - Manually edited lexical data eventually sent back to Basque WordNet and other data providers
- Lindemann & Kliche 2017 (eLex Leiden: this paper)
  - Quantitative and qualitative evaluation of Basque-English BDD
    - Basque WordNet EusWN 3.0, English Princeton WordNet 3.0
    - BabelNet 3.7

Lemma-oriented vs. Concept-oriented

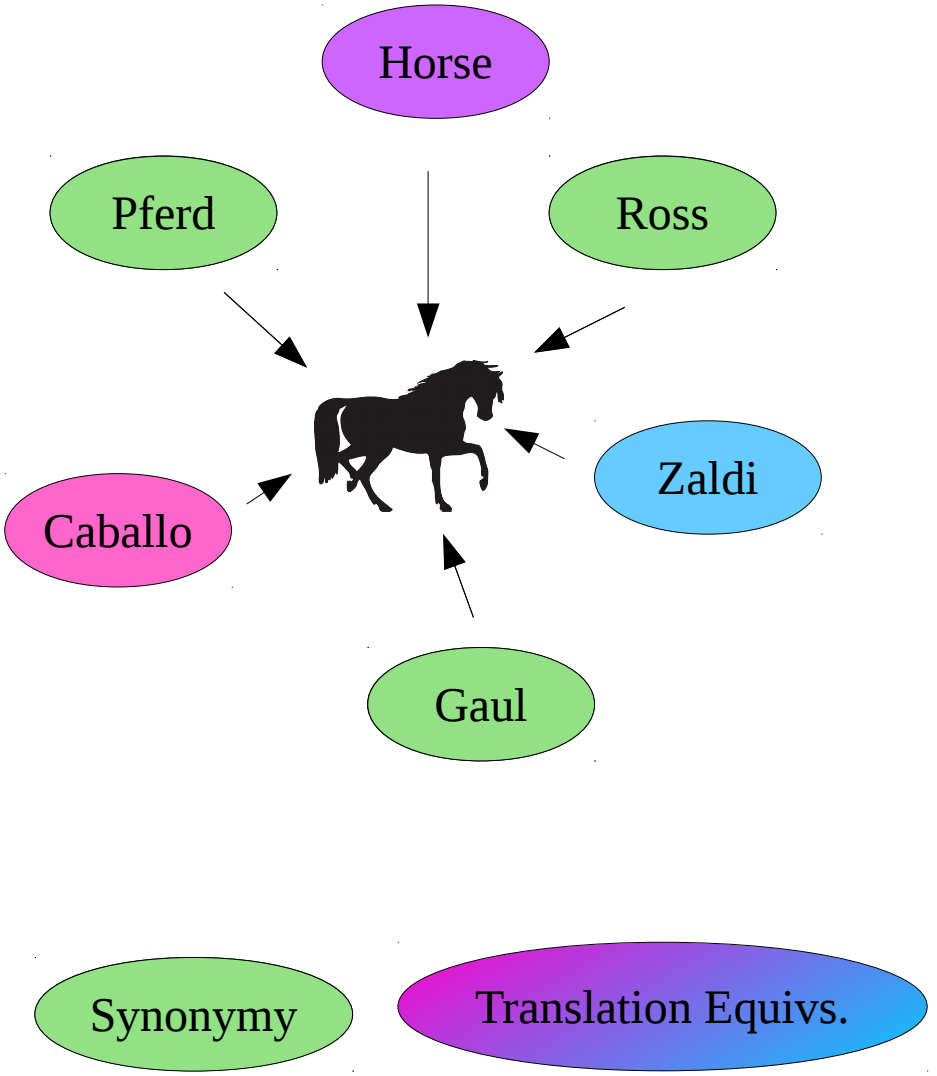


Polysemy: 3 word senses

Lemma-oriented vs. Concept-oriented



Polysemy: 3 word senses



Workflow: A quick walkthrough

- WordNet
  - Download WordNets in table (csv) format:
    - Interlingual Index (Synset IDs)
    - Lexicalisations in the 2 languages
  - Build single XML document

Workflow: A quick walkthrough

- WordNet
  - Download WordNets in table (csv) format:
    - Interlingual Index (Synset IDs)
    - Lexicalisations in the 2 languages
  - Build single XML document

- BabelNet
  - Download complete dump file
  - Retrieve using BabelNet Java API:
    - Synset IDs
    - synset type (concept / NE), English glosses
    - Lexicalisations in the 2 languages, sources
  - Build single XML document

Workflow: A quick walkthrough

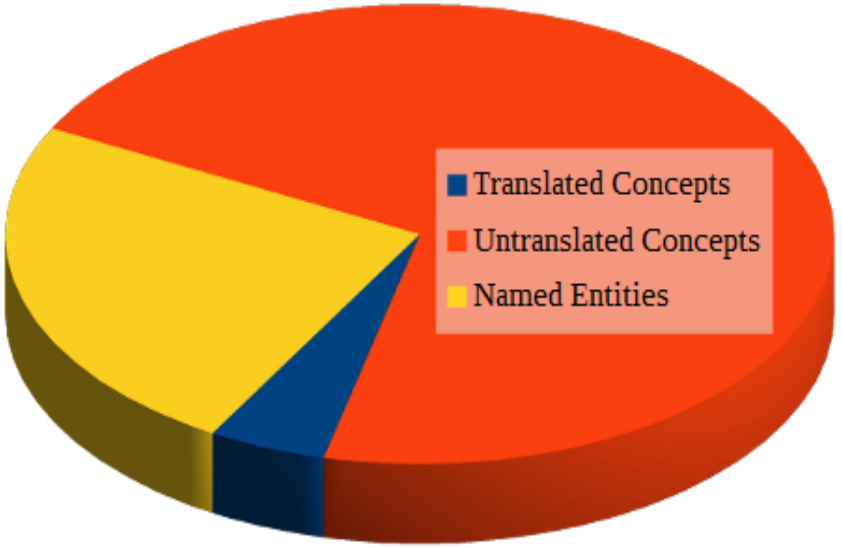
- WordNet
  - Download WordNets in table (csv) format:
    - Interlingual Index (Synset IDs)
    - Lexicalisations in the 2 languages
  - Build single XML document

- BabelNet
  - Download complete dump file
  - Retrieve using BabelNet Java API:
    - Synset IDs
    - synset type (concept / NE), English glosses
    - Lexicalisations in the 2 languages, sources
  - Build single XML document

- Intersection calculations („quantitative evaluation“)
  - Graphical normalization of lemma-strings
    - Initial case, spaces, hyphens
- Assessment of adequacy („qualitative evaluation“)
  - For the evaluators, build a user-friendly view of the XML document
    - Show glosses and lexicalisations
    - Show drop-down menu for choosing assessment value
    - Done using features of *TshwaneLex*

BabelNet 3.7 English-Basque intersection

2.4 Million Synsets



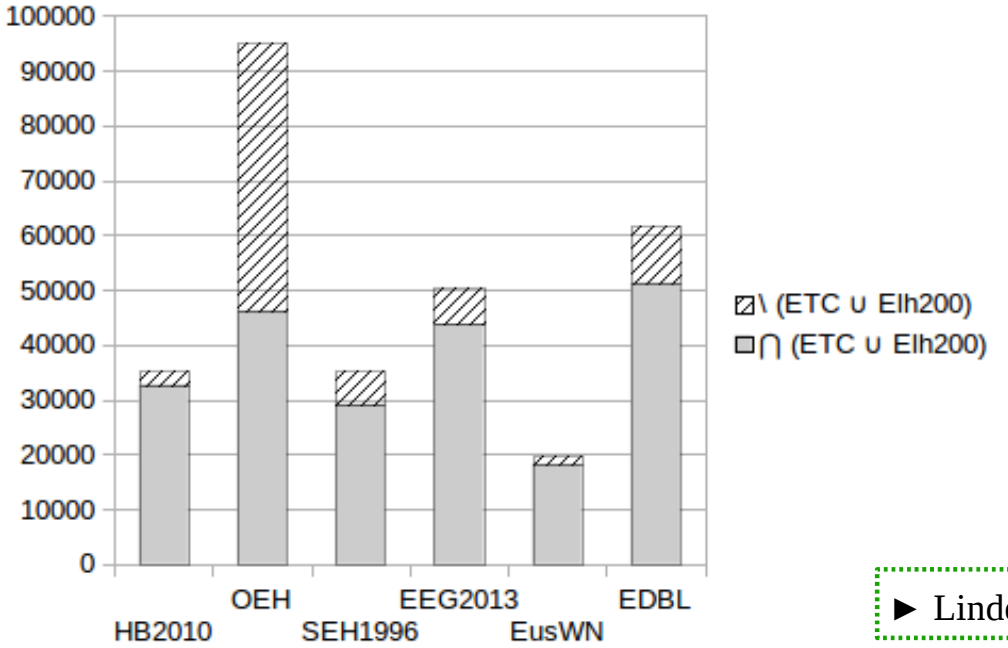
- Named Entities: 24,3 %
  - Place names, proper names
    - May be translated: *Den Haag, The Hague, Haga*
- Untranslated „BabelNet“ Concepts: 71,1%
  - Presumed ‘internationalisms‘
    - *pasta, samba, brahman, yoga, ...*
    - Biology, medicine terms (Greek-Latin)
    - IT terms (English)
    - Abbreviations: *m, cm, kg*
- Translated Concepts: 114,000 (4,6%)
  - 95%+ of what we are looking for belongs to this group

- Sources for Basque Concept translations found in BabelNet:

  Open Multilingual WordNet, Wikidata, Wikipedia Page Titles,
  Wikipedia Redirections, OmegaWiki, Wiktionary, Microsoft Terminology,
  GeoNames, WikiQuotes, WikiQuotes Redirections

Basque lemmata we want to find equivalents for

- Corpus-based frequency headword list for Basque „EusLemStd“:
  58.000 headwords (lemma-signs) that appear both in...
  - ...one of the two very large Basque corpora (20+ occurrences)
    - ETC Hand-selected Basque reference prose corpus (200M tokens, Sarasola, Salaburu & Landa 2013)
    - Elh200 Basque webcorpus (200M tokens, Leturia 2014)
  - ...one of 6 major lexical resources for Basque (4 dictionaries, 2 NLP resources)
- No named entities (proper names, place names)



► Lindemann & San Vicente (2015)

Bilingual Dictionary Draft: Quantitative evaluation

Headwords: intersecting sets		
EusLemStd Basque lemma list	57,919	(100.0%)
EusLemStd ∩ EusWN	18,122	(31.3%)
EusLemStd ∩ EusWN ∩ BabelNet	18,004	(31.0%)
EusLemStd ∩ BabelNet	23,194	(40.0%)

Concepts: intersecting sets	**Noun synsets**	**Verb synsets**	**Adjective synsets**	**Adverb synsets**	**Synsets**
EusWN ∩ EusLemStd	21,533	2,894	106	0	24,533
BabelNet ∩ EusLemStd	31,028	2,914	293	25	34,260

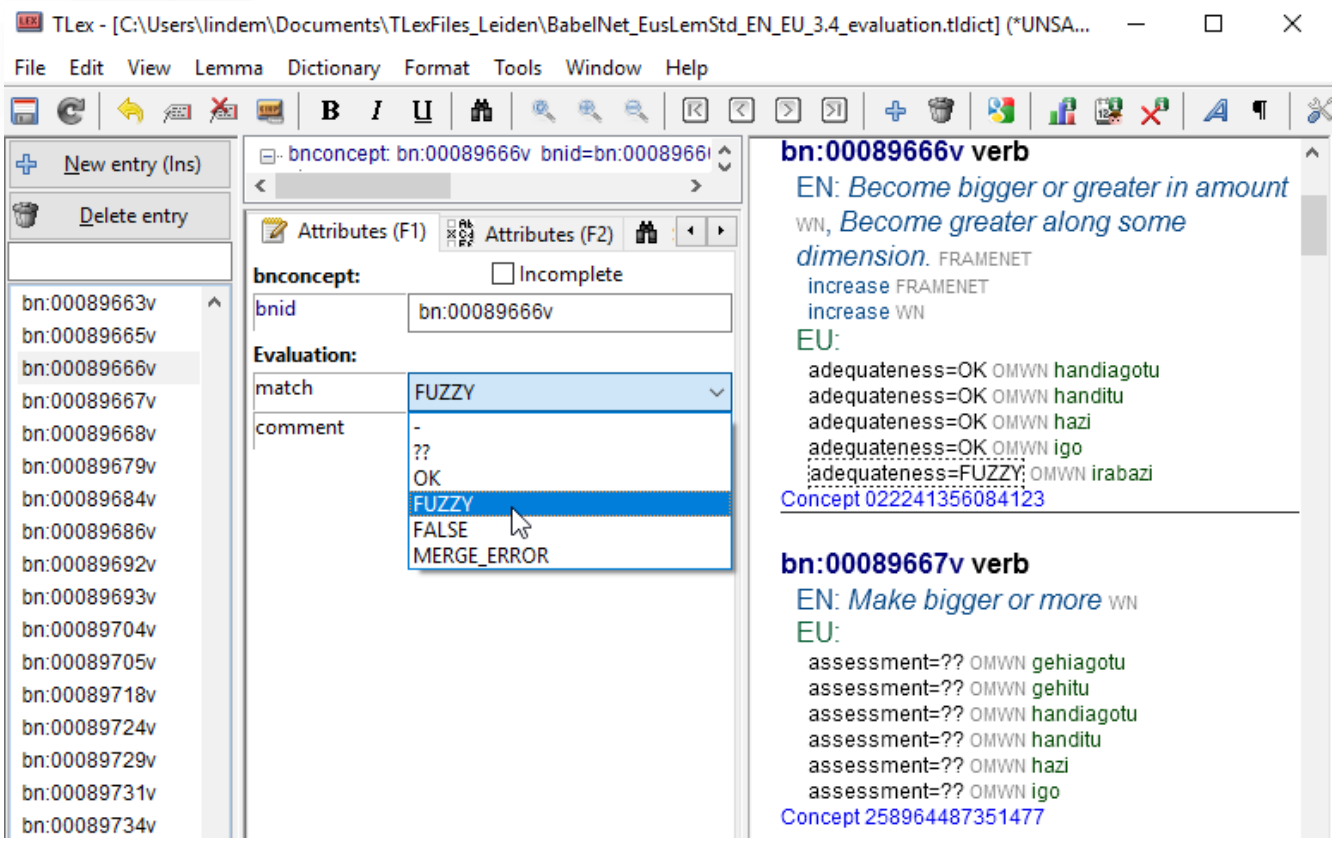
Qualitative evaluation: Manual assessment

Translation equivalents:

- **OK**: correct mapping
- **FUZZY**:
  not false, but without editing not suitable as translation equivalent in a dictionary.
- **FALSE**:
  incorrect mapping

(cf. Fišer, Gantar & Krek 2012, Lindemann et al. 2014)

- **MERGE ERROR**:
  In BabelNet, incorrect merging of concepts



Screenshot: Manual assessments in *TshwaneLex*

Qualitative evaluation: Manual assessment

Translation equivalents:

- **OK**: correct mapping
- **FUZZY**:
  not false, but without editing not suitable as translation equivalent in a dictionary.
- **FALSE**:
  incorrect mapping

(cf. Fišer, Gantar & Krek 2012, Lindemann et al. 2014)

- **MERGE ERROR**:
  In BabelNet, incorrect merging of concepts

Examples:

- **OK**:

  Advanced in years 'aged, elderly, older, senior' – **adindun**, **adineko**, **edadetu**

- **FUZZY**:

  First in order of birth 'firstborn, eldest' – **zahar**
  [the 'autohyponymy' problem, cf. Pociello et al. 2001]

- **FALSE**:

  Provide with a gift 'treat' – **hartu**, **hitz egin**, **tratatu**
  [mismatch to most common sense of 'treat']

- **MERGE ERROR**:

  'Tube, metro, **underground**' (*The London Underground*)
  'Resistance, **underground**' (*A secret group organized to overthrow the government*)

Qualitative Evaluation: Results for WordNet

EusWN/PWN equivalences	Nouns	Verbs	Adjectives	All POS
Total synsets EusWN ∩ EusLemStd	21,533	2894	106	21,533
Monosemous	6,058	201	11	6,270
Polysemous	15,343	2,693	95	18,131
Synsets evaluated	100	100	100	300
Monosemous	50	50	16	
Polysemous	50	50	84	
Synsets all items OK	**87%**	**75%**	94 (**94%**)	**85%**
Monosemous	45 (90%)	37 (74%)		
Polysemous	42 (84%)	38 (76%)		
Synsets OK/FUZZY	**98%**	**94%**	96 (**96%**)	**96%**
Monosemous	49 (98%)	48 (96%)		
Polysemous	49 (98%)	46 (92%)		
Synsets 1+ FALSE	**2%**	**7%**	4 (**4%**)	**4%**
Monosemous	1 (2%)	2 (4%)		
Polysemous	1 (2%)	5 (10%)		

Qualitative Evaluation: Results for WordNet

EusWN/PWN equivalences	Nouns	Verbs	Adjectives	All POS
Total synsets EusWN ∩ EusLemStd	21,533	2894	106	21,533
Monosemous	6,058	201	11	6,270
Polysemous	15,343	2,693	95	18,131
Synsets evaluated	100	100	100	300
Monosemous	50	50	16	
Polysemous	50	50	84	
Synsets all items OK	**87%**	**75%**	94 (**94%**)	**85%**
Monosemous	45 (90%)	37 (74%)		
Polysemous	42 (84%)	38 (76%)		
Synsets OK/FUZZY	**98%**	**94%**	96 (**96%**)	**96%**
Monosemous	49 (98%)	48 (96%)		
Polysemous	49 (98%)	46 (92%)		
Synsets 1+ FALSE	**2%**	**7%**	4 (**4%**)	**4%**
Monosemous	1 (2%)	2 (4%)		
Polysemous	1 (2%)	5 (10%)		

Qualitative Evaluation: Results for BabelNet

BabelNet 3.7	Nouns	Verbs	Adject.	Adverbs	Total
Assessed synsets	200	200	200	25	625
All items OK	179 (89.5%)	163 (81.5%)	188 (94.0%)	23 (92,0%)	553 (**88.5%**)
1+ items OK, and 1+ items FUZZY	3 (1.5%)	14 (7.0%)	2 (1.0%)	0 (0.0%)	19 (**3.0%**)
1+ items OK, and 1+ items FALSE	2 (1.0%)	3 (1.5%)	0 (0.0%)	0 (0.0%)	5 (0.8%)
All items FUZZY	5 (2.5%)	9 (5.5%)	8 (2.0%)	0 (0.0%)	22 (**3.5**%)
1+ items FUZZY, and 1+ items FALSE	1 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
All items FALSE	5 (2.5%)	8 (4.0%)	1 (0.5%)	2 (8.0%)	16 (2.6%)
MERGE_ERROR	5 (2.5%)	3 (1.5%)	1 (0.5%)	0 (0.0%)	9 (1.4%)

Qualitative Evaluation: Results for BabelNet

BabelNet 3.7	Nouns	Verbs	Adject.	Adverbs	Total
Assessed synsets	200	200	200	25	625
All items OK	179 (89.5%)	163 (81.5%)	188 (94.0%)	23 (92,0%)	553 (**88.5**%)
1+ items OK, and 1+ items FUZZY	3 (1.5%)	14 (7.0%)	2 (1.0%)	0 (0.0%)	19 (**3.0**%)
1+ items OK, and 1+ items FALSE	2 (1.0%)	3 (1.5%)	0 (0.0%)	0 (0.0%)	5 (0.8%)
All items FUZZY	5 (2.5%)	9 (5.5%)	8 (2.0%)	0 (0.0%)	22 (**3.5%**)
1+ items FUZZY, and 1+ items FALSE	1 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
All items FALSE	5 (2.5%)	8 (4.0%)	1 (0.5%)	2 (8.0%)	16 (2.6%)
MERGE_ERROR	5 (2.5%)	3 (1.5%)	1 (0.5%)	0 (0.0%)	9 (1.4%)

Qualitative Evaluation: BabelNet sources

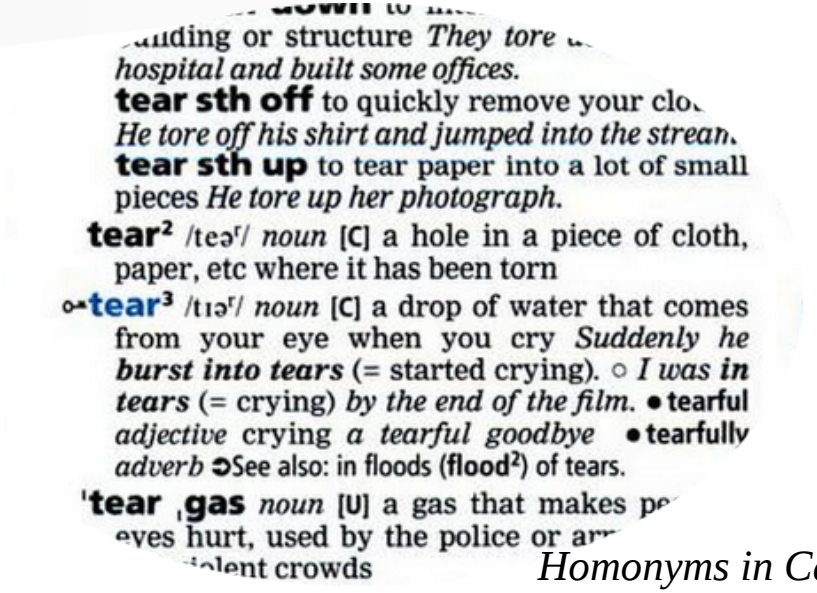
BabelNet 3.7	OK	FUZZY	FALSE	MERGE ERROR	(Asses-ments)
All Sources	*1,211 (88.9%)*	*63 (4.6%)*	*44 (3.2%)*	*44 (3.2%)*	*1,362*
Open Multilingual WordNet	717 (89.2%)	49 (6.1%)	28 (3.5%)	10 (1.2%)	804
Wikidata	57 (93.4%)	0 (0.0%)	1 (1.6%)	3 (4.9%)	61
Wikipedia	194 (87.8%)	5 (2.3%)	6 (2.7%)	16 (7.2%)	221
BabelNet	3 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3
Wikipedia Redirections	13 (52.0%)	3 (12.0%)	4 (16.0%)	5 (20.0%)	25
OmegaWiki	75 (91.5%)	2 (2.4%)	0 (0.0%)	5 (6.1%)	82
Wiktionary	132 (92.3%)	4 (2.8%)	5 (3.5%)	2 (1.4%)	143
Microsoft Terminology	20 (87.0%)	0 (0.0%)	0 (0.0%)	3 (13.0%)	23

Qualitative Evaluation: BabelNet sources

BabelNet 3.7	OK	FUZZY	FALSE	MERGE ERROR	(Asses-ments)
All Sources	*1,211 (88.9%)*	*63 (4.6%)*	*44 (3.2%)*	*44 (3.2%)*	*1,362*
Open Multilingual WordNet	717 (89.2%)	49 (6.1%)	28 (3.5%)	10 (1.2%)	804
Wikidata	57 (93.4%)	0 (0.0%)	1 (1.6%)	3 (4.9%)	61
Wikipedia	194 (87.8%)	5 (2.3%)	6 (2.7%)	16 (7.2%)	221
BabelNet	3 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3
Wikipedia Redirections	13 (52.0%)	3 (12.0%)	4 (16.0%)	5 (20.0%)	25
OmegaWiki	75 (91.5%)	2 (2.4%)	0 (0.0%)	5 (6.1%)	82
Wiktionary	132 (92.3%)	4 (2.8%)	5 (3.5%)	2 (1.4%)	143
Microsoft Terminology	20 (87.0%)	0 (0.0%)	0 (0.0%)	3 (13.0%)	23

Post-processing / editing: Some central issues

- Creation of a headword-oriented dictionary
  - Transformation of XML containing dictionary draft
  - Homonym disambiguation
    - In WordNet, Wikipedia, BabelNet, homonymy = polysemy
    - In a dictionary, homonymy ≠ polysemy



…ilding or structure *They tore …*
hospital and built some offices.
tear sth off to quickly remove your clo…
He tore off his shirt and jumped into the strean.
tear sth up to tear paper into a lot of small pieces *He tore up her photograph.*
tear[2] /teə^{r}/ *noun* [C] a hole in a piece of cloth, paper, etc where it has been torn
tear[3] /tɪə^{r}/ *noun* [C] a drop of water that comes from your eye when you cry *Suddenly he* ***burst into tears*** (= started crying). ○ *I was* ***in tears*** (= crying) *by the end of the film.* ● **tearful** *adjective* crying *a tearful goodbye* ● **tearfully** *adverb* ➲See also: in floods (**flood**[2]) of tears.
'tear ˌgas *noun* [U] a gas that makes pe… eyes hurt, used by the police or ar… …olent crowds

Homonyms in Cambridge Learner's Dictionary with CD-ROM, 2007 [img source]

- Representation of polysemy
  - Does the draft entry contain all word senses?
  - Is the splitting of senses...
    - ...too fine-grained?
    - ...even redundant?
    - ...too coarse-grained?
- Other issues: cf. Benjamin 2016
- Restrictive licensing of some WordNets

WN/BN bootstrapping for EUS-EN: Result Overview

- Recall on initial Basque headword list
  - EusWN / PWN alone 30%
  - BabelNet 40%
- Precision
  - EusWN / PWN alone 90%
  - BabelNet 90%

- BabelNet
  - Higher Recall than WN alone
  - Similar Precision

WN/BN bootstrapping for EUS-EN: Result Overview

- Recall on initial Basque headword list
  - EusWN / PWN alone 30%
  - BabelNet 40%
- Precision
  - EusWN / PWN alone 90%
  - BabelNet 90%

- BabelNet
  - Higher Recall than WN alone
  - Similar Precision

Does this approach work with language pairs 'un-resourced' in Bilingual Lexicography?

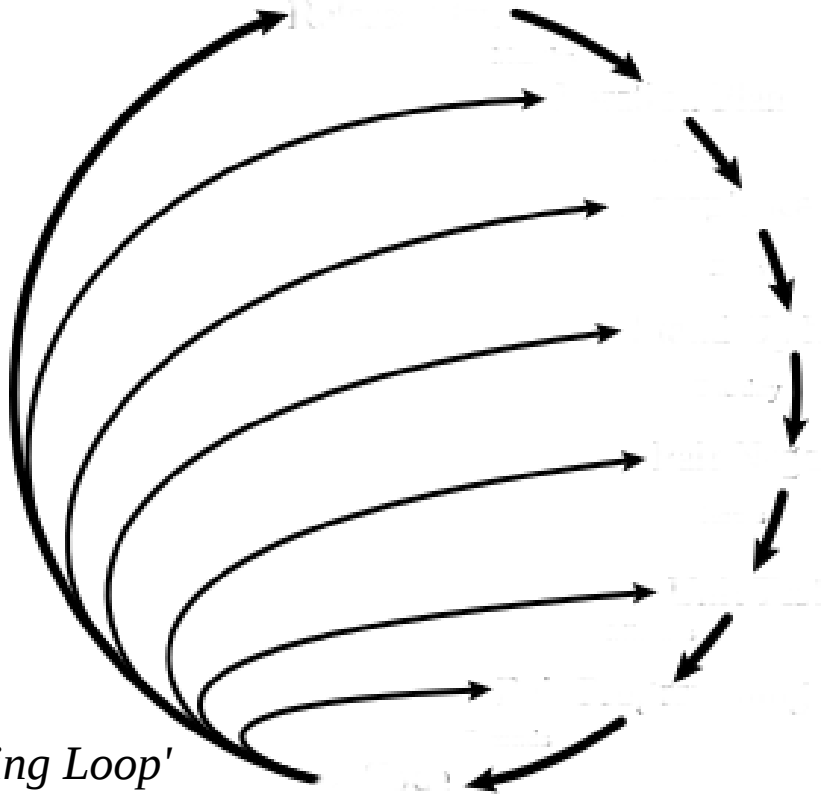
YES, it does

Application example: WordNet/BabelNet bootstrapping for EUS-SLO

- Basque (EUS) – Slovene (SLO): A totally uncovered pair of 'smaller' languages
- Quantitative Evaluation
  - Recall: Synsets that contain 1+ Basque standard headword and 1+ Slovene item
    - EusWN / SloWNet 20% (66% of 30%)
    - BabelNet 31% (78% of 40%)
  - Recall on 5,000 most frequent Basque headwords (BabelNet): 74% (3,707)
  - Recall on 20,000 most frequent Basque headwords (BabelNet): 53% (10,549)
- Qualitative Evaluation
  - Precision: Unknown. EN-SL precision to be measured first.

Conclusions and further work

- Bilingual Dictionary Draft for Basque-English including sense-to-sense mappings
  - Encouraging recall and precision rates; can be applied to other language pairs
- Preliminaries for a research project
  - Bilingual Dictionary Drafts for many uncovered language pairs
  - Data model that allows
    - Manual and semi-automated (bulk) editing
    - Edition of e-dictionaries including more item types
    - Retro-updating of original resources: 'Bootstrapping Loop'
  - Engagement of lexicographers for editing 'their' language pair
  - Edition of a new series of bilingual dictionaries with Basque



'Bootstrapping Loop'

Image Source: Wikimedia Commons

The research leading to these results has received funding from the Basque Government (Research Group IT665-13). Funding is gratefully acknowledged.

Thank you for your attention
Eskerrik asko, bedankt!

david.lindemann@ehu.eus
fritz.kliche@uni-hildesheim.de

References

- Benjamin, M. (2016). Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary. In Proceedings of the Eighth Global WordNet Conference (pp. 27–33). Bucharest: Alexandru Ioan Cuza University of Iasi.
- Bond, F., & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In Proceedings of the The 51st Annual Meeting of the Association for Computational Linguistics (pp. 1352–1362).
- Fišer, D., Gantar, P., & Krek, S. (2012). Using explicitly and implicitly encoded semantic relations to map Slovene Wordnet and Slovene Lexical Database. In Semantic Relations-II. Enhancing Resources and Applications Workshop Programme (p. 77).
- Gale, W. A., & Church, K. W. (1991). Identifying Word Correspondences in Parallel Texts. In Proceedings of the ACL Workshop on Speech and Natural Language (pp. 152–157). Stroudsburg, PA: Association for Computational Linguistics.
- Héja, E. (2010). The Role of Parallel Corpora in Bilingual Lexicography. In N. Calzolari, K. Choukry, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), Proceedings of LREC 2010. Valetta.
- Kazakov, D., & Shahid, A. R. (2013). Using Parallel Corpora for Word Sense Disambiguation. (pp. 336–341). Proceedings of RANLP 2013, Hissar.
- Lefever, E. (2012). ParaSense: parallel corpora for word sense disambiguation (PhD Thesis). Universiteit Gent, Gent.
- Lindemann, D., & San Vicente, I. (2015). Building Corpus-based Frequency Lemma Lists. Procedia - Social and Behavioral Sciences, 198, 266–277.
- Lindemann, D., & San Vicente, I. (2016). Bilingual Dictionary Drafting: Connecting Basque word senses to multilingual equivalents. In Proceedings of EURALEX 2016 (pp. 898–905). Tbilisi.
- Lindemann, D., Saralegi, X., San Vicente, I., Manterola, I., & Nazar, R. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density language pair. In Proceedings of EURALEX 2012 (pp. 563–576). Bolzano.
- Navigli, R., & Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 216–225). Stroudsburg.
- Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2), pp. 121–142.
- Saralegi, X., Manterola, I., & San Vicente, I. (2012). Building a Basque-Chinese Dictionary by Using English as Pivot. In Proceedings of LREC 2012. Istanbul.
- Speer, R., & Havasi, C. (2012). Representing General Relational Knowledge in ConceptNet 5. In Proceedings of LREC 2012. Istanbul.
- Varga, I., Yokoyama, S., & Hashimoto, C. (2009). Dictionary generation for less-frequent language pairs using WordNet. Literary and Linguistic Computing, 24(4), 449–466.