CILC 2015
7.º Congreso Internacional de Lingüística del Corpus

aelinco
Spanish Association for Corpus Linguistics

Building corpus-based frequency lemma lists

...a corpus-based frequency lemma list for Basque

Universidad del País Vasco
Euskal Herriko Unibertsitatea

David Lindemann
UPV-EHU University of the Basque Country
david.lindemann@ehu.es

Iñaki San Vicente
Elhuyar Foundation
i.sanvicente@elhuyar.com

elhuyar
Hizkuntza eta Teknologia

Overview

- Introduction
  - Lemma Frequency in Lexicography
  - Motivation for this Study
  - The State of the Art
- Experiments
  - Resources: Basque Corpora
  - Resources: Basque Dictionary Lemma Lists
  - Frequency Lemma List Processing
- Results
  - Comparison of Frequency Lemma Lists
  - Dictionary Lemma Lists in the Corpora
  - A corpus-based frequency lemma list for EuDeLex
- Conclusions and Future Work

Lemma Frequency and Lexicography

Reasons for taking frequency as basic criterion for building a lemma list

- Lemma frequency and headword look-up frequency are related
  - de Schryver et al. 2010: Clear relation in the “top few thousand”
  - Wolfer et al. 2014: Clear relation, analysing *de.wiktionary* log files
- Frequency data is useful information for a lexicographer
- Frequency data may be included in the dictionary article and passed on to the dictionary user
  - Kilgarriff 1997

Motivation for this study

EuDeLex

ehk EUSKAL HIZKUNTZA ETA KOMUNIKAZIOA SAILA UPV/EHU | www.ehk.ehu.es

Universidad del País Vasco Euskal Herriko Unibertsitatea

Wörterbuch Deutsch-Baskisch - Alemaniera-Euskara Hiztegia

- To define a Basque lemma list for a first edition of the Basque→German part of **EuDeLex**, a German-Basque bilingual dictionary currently being developed at UPV-EHU
- To survey and prove criteria for including headword candidates in a dictionary lemma list, and to propose a methodology for
  - Future Bilingual Dictionary projects with Basque
  - Future Bilingual Dictionary projects for other less resourced language pairs
- This publication: To provide guidelines comprehensive to (beginning) lexicographers

The State of the Art: DeReWo corpus-based lemma list

- DeReWo: Frequency word and lemma lists for German
  - based on the DeReKo corpora (Kupietz et al. 2010)
- Lemma list building method (IDS 2009)
  - Intersections of the frequency lemma list extracted from corpora with previously existing dictionary lemma lists taken as accurate headword candidates;
  - The remaining list entries, that don't appear in previously published dictionaries, are classified by semi-automatic methods (that is, in groups) or by hand as accurate headword candidates (or not).
- Practical experience with DeReWo (editing EuDeLex)
  - After manual editing of a 10% sample, DeReWo list and dictionary lemma list are 95% the same (Lindemann 2014)

The State of the Art: Basque frequency dictionaries

- Sarasola 1982
  - Based on 1977 corpus
  - 3.000 lemmata
- Etxebarria & Mujika 1987
  - Based on voice recording and radio broadcast corpus
  - 3.154 lemmata
- UZEI 2004
  - Based on 4.5 million token 20th century (literature) corpus

Resources: Basque Corpora

- ETC (UPV-EHU: Sarasola et al. 2013)
  - Basque press, literature, science and television broadcast texts selected by hand, and the Basque *Wikipedia*
  - 205 million tokens
- Elhuyar Web-Corpus Elh124 (Leturia 2012)
  - Built from the web by the *seed-word* method
  - 124 million tokens
- Elhuyar Web-Corpus Elh200 (Leturia 2014)
  - Built from the web by the *crawling* method
  - 200 million tokens

Resources: Basque Dictionaries / Lexical Data Bases

- *Hiztegi Batua* (HB) (Euskaltzaindia 2008)
  - Language Academy. 35.640 headwords (homographs counted once).
- *Orotariko Euskal Hiztega* (OEH) (Mitxelena & Sarasola 1988)
  - Corpus-based. 89.296 headwords, 36.676 subentries (homographs counted once).
- Elhuyar EU-ES (ElhDic) (Elhuyar 2013)
  - 64.459 Basque headwords (homographs counted once)
- Basque WordNet (EusWN) (Pociello 2007; Pociello et al. 2011)
  - Equivalents to English PWN synsets. 26.886 lexical units
- *Euskararen Datu Base Lexikala* (EDBL) (Aldezabal et al. 2001)
  - Designed for NLP purposes. 64.737 canonical forms (lemmata)

Frequency Lemma List Processing (1)

- **ETC**
  - Lemma list with frequency data
  - 47.498 lemma signs (without POS-disambiguation, min.freq 10)
- **Elhuyar Web-Corpora**
  - Lemmatizing and POS-tagging with *Eustagger* (Aduriz et al. 1996)
  - Frequency data for three granularity levels
    - Homograph lemmata (without POS-disambiguation)
      - 79.129 lemma signs, min.freq 20
    - Main POS-category (*noun, verb, adjective...*)
      - 74.617 lemma signs, min.freq 40
    - POS-subcategories (*noun, proper noun, place name...*)
      - 59.757 lemma signs, min.freq 40

Frequency Lemma List Processing (2)

- Processing steps
  - Pre-processing: Graphical normalization
    - normalization to lower case
    - both spaces and hyphens between n-gram elements > “_”
    - end-of-word “-” hyphens cut
  - Lemmatizing, syntactical tagging
  - Calculation of relative frequency values (percentages)
- For this experiment (regarding definition of headword candidates)
  - Only single-word lexical units (unigrams)
  - Lemma frequency data
    - PoS disambiguated (2 granularity levels)
    - without regarding homography (same PoS) and polysemy

Frequency Lemma List Processing (3)

- 3 granularity levels: NoPOS, POS, POS_POS2
- Example: *alegia* in Elh200 corpus

Level	*Rank*	*Occurrences*	*POS*
NoPOS			
	539	46237	(lemma-sign)
POS			
	609	41106	conjunction
	3378	5129	noun
	~~1475870~~	~~1~~	~~adverb~~
	~~1507920~~	~~1~~	~~adjective~~
POS_POS2			
	618	41106	conjunction
	3882	4208	common noun
	10407	921	place name
	~~1440023~~	~~1~~	~~adjective~~
	~~1880968~~	~~1~~	~~adverb~~

evidence for dictionary macrostructure building: lemma-sign to be included

evidence for basic distribution of dictionary entry content: syntactical entities

Accidental false taggings? Correct hapax taggings?
Minimum frequency threshold for lexicography: 20 occurrences (Sinclair 2005)

Frequency Lemma List Processing (4)

- Typology of Noise (Frequency lemma list entries that are unaccurate headword candidates)
  - Words from other languages that are homograph to Basque lemmata
    - contaminates the frequency data for the real Basque lemma
    - example *el*
  - Words from other languages that are not homograph to any Basque lemma
    - example *con*
  - Errors in automatic linguistic tagging
    - wrong lemmatisations of Basque words
    - wrong PoS-tagging of Basque words

Frequency Lemma List Processing (5)

- Workaround strategies for noisy results (list entries)
  1. If *Eustagger* throughout the corpus has used a high number of different POS-tags for the same lemma, to suppress that lemma;
  2. To suppress a lemma if it appears to have a high frequency rank on our list, but counts with no usage example in the corpus-based OEH dictionary, as it has been proposed before to use the amount of usage examples in OEH as indicator for frequency (Sarasola et al. 2008);
  3. To suppress a lemma if its *rfreq* value appears to be very different for the two processed big web-corpora.
- Result: No satisfactory clean-up
- Conclusions for the time being:
  - Frequency data alone is not enough to deal with this problem
  - The planned manual editing process will show the overall accuracy

Results (1): Comparison of frequency lemma lists I

- Spearman Rank Correlation (Kilgarriff 2001)

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82	1				
UZEI04	0.5542798107	1			
ETC	0.4056939783	0.4641690074	1		
Elh124	0.4150037011	0.5536959952	0.5627681191	1	
Elh200	0.3805365731	0.4675802057	0.5265724492	**0.9009107076**	1

$$\rho = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2-1)}$$

Computes the distance between two rankings by comparing the positions *n* (500) elements have in one ranking and the other

Results (2): Comparison of frequency lemma lists II

- Coverage (Baroni et al. 2009)

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82		26.13%	7.11%	4.13%	3.36%
UZEI04	91.76%		22.14%	14.13%	11.51%
ETC	93.78%	83.12%		41.58%	34.63%
Elh124	95.38%	93.04%	72.91%		89.80%
Elh200	95.52%	93.08%	74.61%	91.70%	

$$coverage(C_a/C_b) = \frac{Na \cap Nb}{Na}$$

Measures how many of the lemmata that reach a minimum frequency *t* (*t*=20) in C_b occur also at least *t* times in C_a

Results (3): Comparison of frequency lemma lists III

- Enrichment (Baroni et al. 2009)

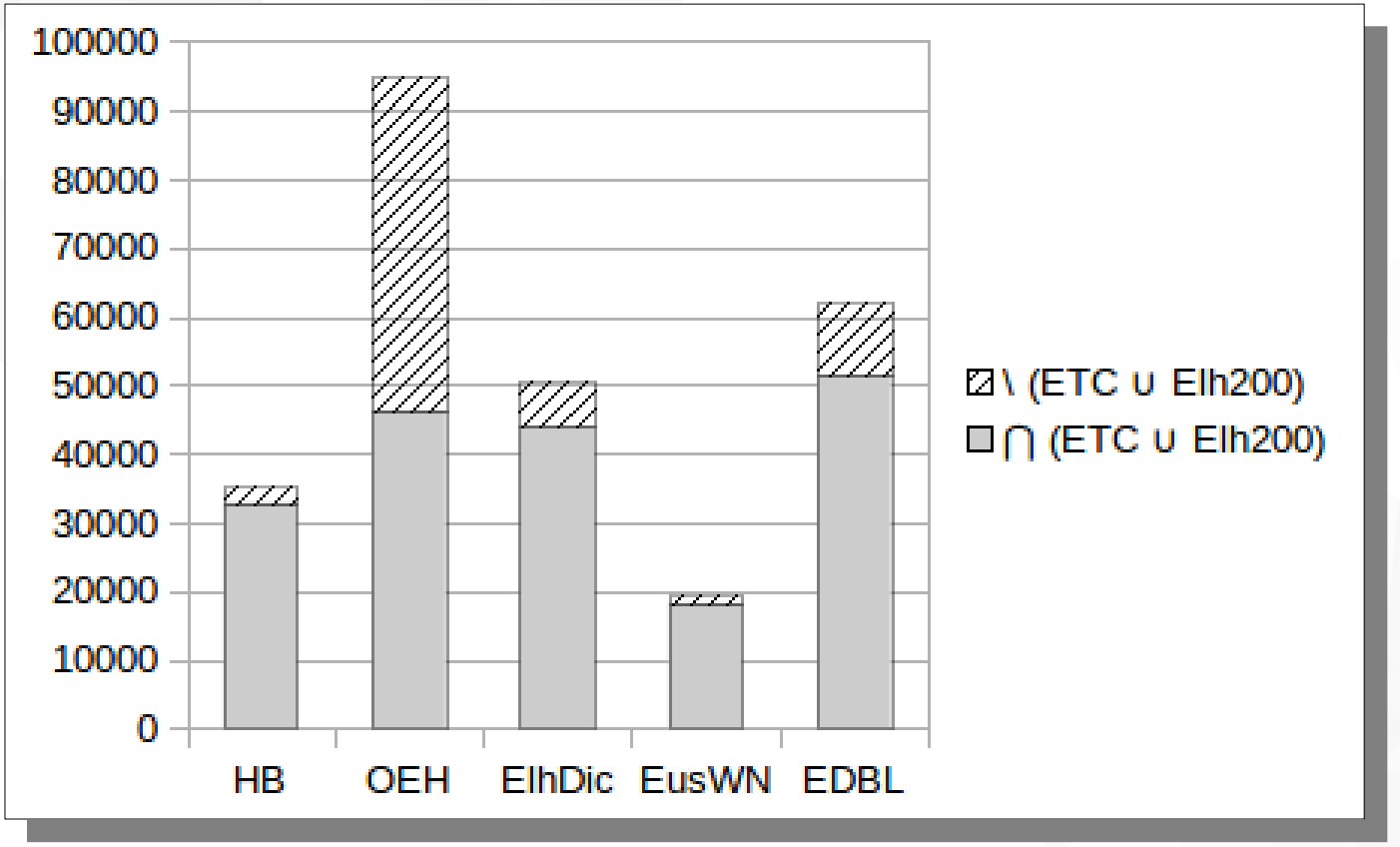
	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82		0.60%	0.20%	0.02%	0.01%
UZEI04	80.41%		0.64%	0.16%	0.14%
ETC	94.85%	66.08%		5.39%	7.83%
Elh124	93.81%	82.24%	18.78%		10.84%
Elh200	94.85%	82.65%	21.82%	28.26%	

$$enrichment(C_a/C_b) = \frac{Nb}{Sa}$$

Measures the proportion of lemmata that occur less than *t* (*t*=20) times in C_a but more than *t* times C_b, with respect to the total number of lemmata below the frequency threshold *t* in C_a. This would describe which is the proportion of lemmata for which C_b can provide information but C_a can not.

Results (4): Dictionary headwords in the Corpora

- Intersections and Complement Sets of Basque Dictionary Lemma Lists and Corpus-based Frequency Lemma Lists (unigrams only)



Results (6): A basic lemma list for EuDeLex

- Intersection of our frequency lists (ETC ∪ Elh200) with existing Basque lex. resources
  - 75.481 headwords
- Intersection of our frequency lists with EDBL
  - 57.919 headwords (20+ occurrences)
- Reasons for building a first edition upon EDBL
  - Relatively, the biggest intersection of handmade lemma list and corpus lemmata
  - Use of EDBL syntactical tags for subdividing the dictionary article in syntactic entities
  - Dictionary editing by hand: detection of gaps and possible errors in EDBL, feedback to EDBL developers

```
<homograph homograph="alegia">
    <syntactical_entity lemma="alegia" pos="conjunction" corpus_counts="41.106">
        <sense equivalent="that's to say"/>
    </syntactical_entity>
    <syntactical_entity lemma="alegia" pos="noun_common_noun" corpus_counts="4.208">
        <sense equivalent="allegory"/>
    </syntactical_entity>
    <syntactical_entity lemma= "Alegia" pos="noun_toponym" corpus_counts="921">
        <sense equivalent="Alegia" Explain="City in Gipuzkoa"/>
    </syntactical_entity>
</homograph>
```

xml entry model based on EDBL syntactical tags (simplified sense groups)

Conclusions

- Built a corpus-based frequency lemma list for Basque
- Proposed a basic macrostructure for a new dictionary
- Proposed a division of the microstructure in syntactic entities
- Provided frequency data for all of these entities

Future Work

- Inclusion of Multi-Word Lexical Units (n-grams)
- Same set of experiments for diachronic, regional, domain etc. corpora
- Basque Meta-Dictionary („Dictionary of Dictionaries“, „Lexicographical corpus“)

Thank you for your attention
Eskerrik asko, muchas gracias

david.lindemann@ehu.eus
i.sanvicente@elhuyar.com