Multilingual lexicography for lesser resourced languages: The case of Basque

Euskara Institutua
UPV EHU
University of the Basque Country

David Lindemann
david.lindemann@ehu.eus

Motivation

To produce dictionary drafts for a series of Bilingual Dictionaries not existent so far

State of the Art

Today, we have Bilingual Dictionaries of a significant size with ES, EN, FR, RU

Frequency lemmalist for Basque

Built by comparing lemmata of ETC and Elh200 corpora (200M token each) to the lemmata present in 6 standard Basque dicts. & resources

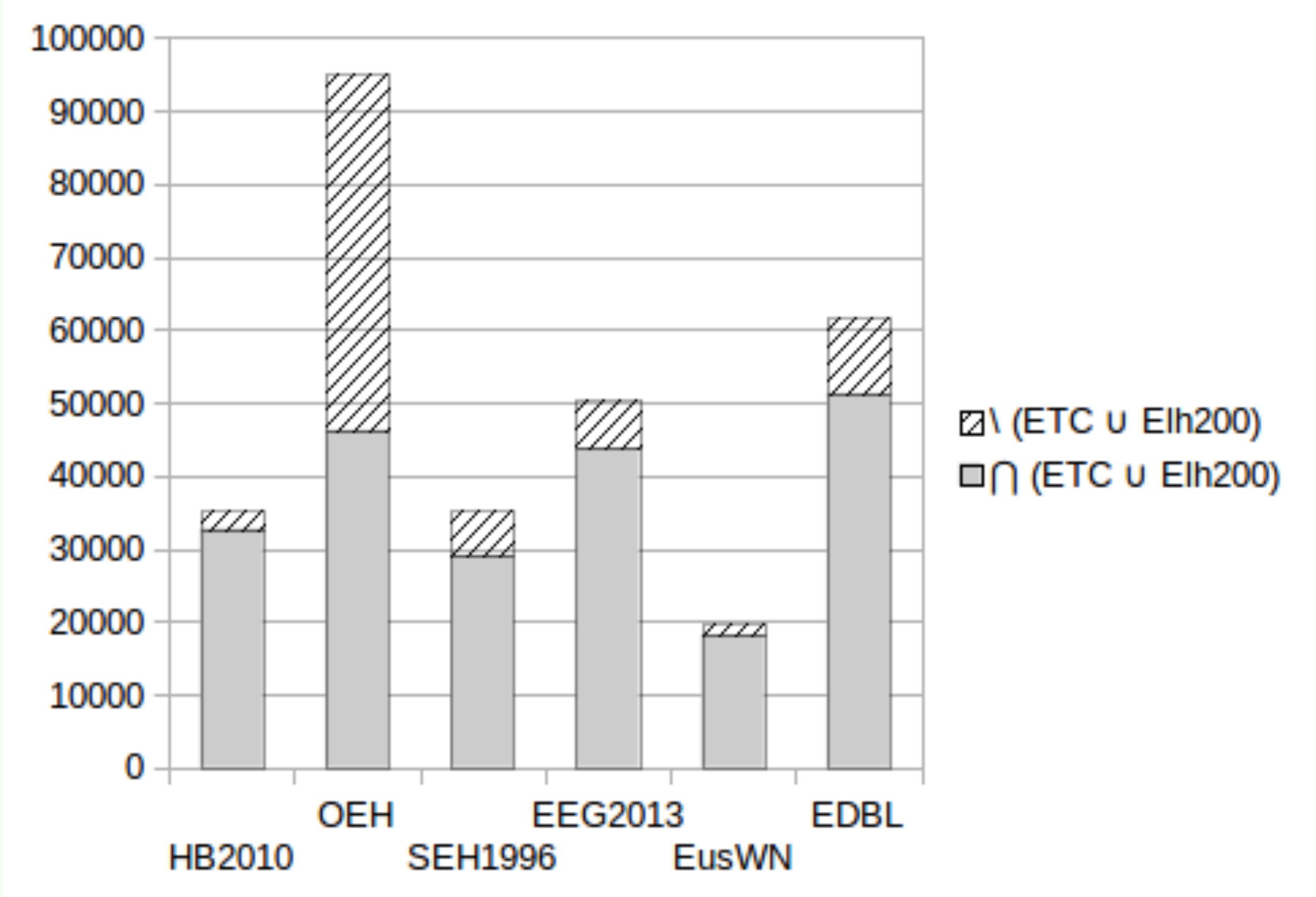


Fig: Basque lemma-signs in the big 2 corpora and 6 hand-made standard resources (Lindemann & San Vicente 2015)

Existing resources for drafting a basic Basque dictionary

* Lexical Database for Basque EDBL (Aldazabal et al. 2001)
* Basque WordNet EusWN (Pociello et al. 2011)
* Basque Dictionary lemmalists (cf. Lindemann & San Vicente 2015)
* Corpus-based frequency lemma list (Lindemann & San Vicente 2015)

Advantages

1) Automatically drafted first version to work on
2) Basic vocabulary present (50.000 lemma-signs)
3) Only lemmata present in corpora (really used words)
4) Frequency data at 3 levels
5) Syntactical entities drafted
6) Revision of the draft by hand: evidence for improving EDBL

Starting point: Lemma-signs present in corpora and **EDBL**:
* Lemma-sign disambiguated as **syntactical entity**
* **Frequency data** for all of these entities

Level	*Rank*	*Occurrences*	*POS*
NoPOS			
	539	46237	(lemma-sign)
POS			
	609	41106	conjunction
	3378	5129	noun
POS_POS2			
	618	41106	conjunction
	3882	4208	common noun
	10407	921	place name

Table: Lemma-sign ***alegia***: Elh200 corpus frequency data and EDBL-based POS-tags on 3 granularity levels

WSD and equivalents drafted with WordNet data

* Not a new approach (cf. EuroWordNet, BabelNet, etc.)
* New: Basque lemma-signs as nodes for linked multilingual lexical data
* Lexical, conceptual gaps in the draft dict to be filled with evidence from Basque

Advantages

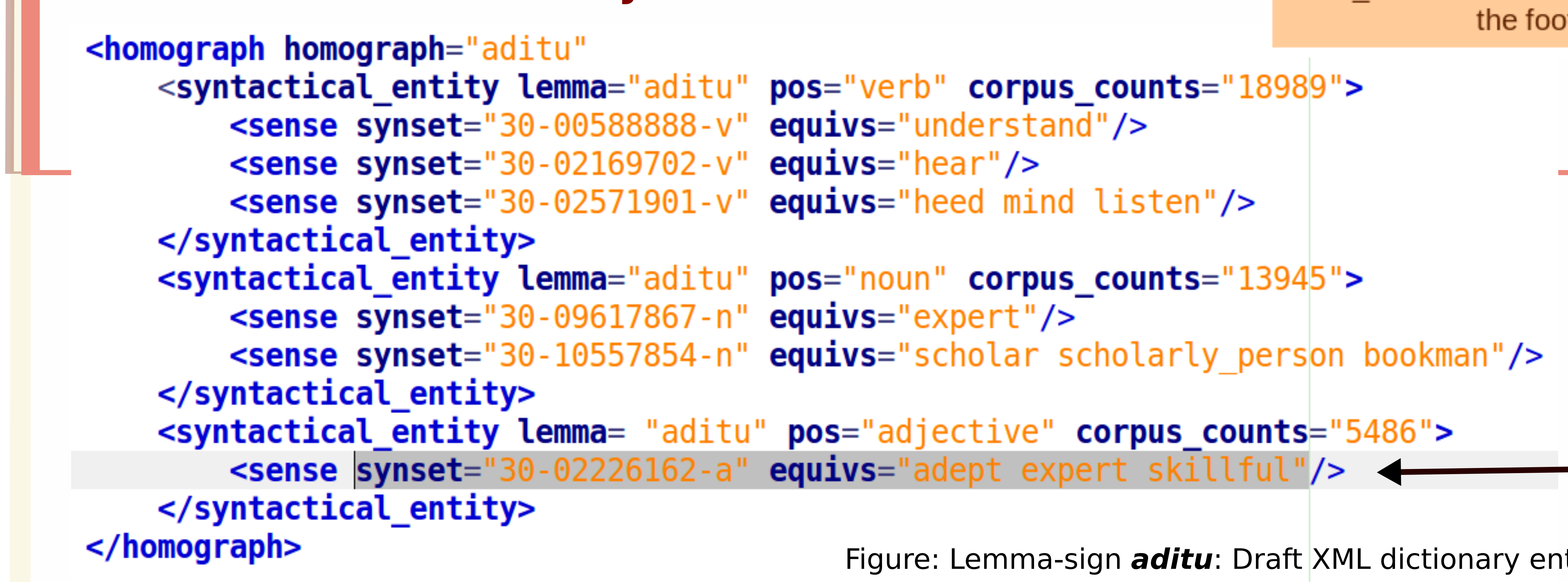
1) Multilingual Dictionary Drafting approach
2) Automatically drafted Basque WSD to work on
3) Basic vocabulary present (18.216 lemma-signs)
5) Only lemmata present in corpora (really used words)
5) Frequency data for syntactical entities present in corpora
6) Translation equivalents, synonyms & more WN semrels
7) Revision by hand: evidence for improving EusWN, and, by chance, WordNet in general

Lexical Unit EU	Definiton EN	EU	EN	CAT
adar_1	one of the bony outgrowths on the heads of certain ungulates	adar_1	horn_2	banya_1
adar_2	a railway line connected to a trunk line	adar_2	branch_line_1 spur_track_1 spur_5	enforcall_1 forcall_1
adar_3	a warning signal that is a loud wailing sound	adar_3 sirena_2 turuta_5	siren_3	
adar_4	a local branch of some fraternity or association	adar_4	chapter_3	capítol_2
adar_5	a division of a stem, or secondary stem arising from the main stem of a plant	adar_5 abar_2 besanga_1 beso_12	branch_2	branca_1 branc_1
adar_6	an alarm device that makes a loud warning sound	sirena_4 adar_6 turuta_6	horn_9	
adar_7	a device used for easing the foot into a shoe	zapata_sartzeko	shoehorn_1	calçador_1

Table: Lemma-sign ***adar***: EusWN senses, linked data from EN, CAT WordNets... and a sense gap

Revision by hand: Sense gap detected and filled

Result: EDBL & EusWN joint datasets



```
<homograph homograph="aditu"
    <syntactical_entity lemma="aditu" pos="verb" corpus_counts="18989">
        <sense synset="30-00588888-v" equivs="understand"/>
        <sense synset="30-02169702-v" equivs="hear"/>
        <sense synset="30-02571901-v" equivs="heed mind listen"/>
    </syntactical_entity>
    <syntactical_entity lemma="aditu" pos="noun" corpus_counts="13945">
        <sense synset="30-09617867-n" equivs="expert"/>
        <sense synset="30-10557854-n" equivs="scholar scholarly_person bookman"/>
    </syntactical_entity>
    <syntactical_entity lemma= "aditu" pos="adjective" corpus_counts="5486">
        <sense synset="30-02226162-a" equivs="adept expert skillful"/>
    </syntactical_entity>
</homograph>
```

Figure: Lemma-sign ***aditu***: Draft XML dictionary entry

Automatically detected sense gap filled by hand

Acknowledgements:
This study has been supported by the project IT665-13, funded by the Basque Government, and by the project EC FP7/SSH-2013-1 AThEME (613465), funded by the European Commission. Funding is gratefully acknowledged.