Eu Delex: Creating a German-Basque Electronic Dictionary

Functions

Primarily targeted user groups of a first edition are (1) Basque-L1 German learners, and (2) German to Basque translators.

We propose a database with an open structure that allows to include more data in the future, and to derive different dictionary macro- and microstructures from it. For design and editing of this database, we use the *TshwaneLex* software (De Schryver & Joffe 2005).

Macrostructure

Frequency-based lemmalists

- It has been proved that the most frequent words are actually the words most frequently looked up by dictionary users; this is true for the top few thousand (De Schryver et al. 2010).
- Frequency data is useful information for both dictionary editor and user.

German

• We define a lemma list based on a corpus-based frequency word list (DeReWo-40.000, IDS 2009), revise it by hand, and include frequency data in the published dictionary. GFL learners' basic vocabulary is fully covered by DeReWo.

Basque

• We develop a Basque frequency-based lemmalist from corpus-based frequency word lists (Elhuyar WebCorpus, ETC UPV-EHU) and the Basque Language Academy's *Hiztegi Batua* (Lindemann & San Vicente, in prep.)

Research and Working Steps

- Proposals for macrostructure and microstructure
- 10% of DE Lemmalist: Edition of Dictionary entries DE>EU
- Scripting for User Interface, publication of a preliminary version
- Bilingual Dictionary Drafting (Lindemann, Manterola, Nazar et al. 2014)
- Result evaluation by hand (Letter A)
- Possible pasting of draft data into EuDeLex database (in prep.)
- Edition and publication of all dictionary entries DE>EU (planned)
- Edition and publication of dictionary entries EU>DE (planned)
 - Drafting by inverting DE>EU articles

German-Basque Corpora

• See Lindemann (2013), Lindemann, Manterola, Nazar et al. (2014)

Parallel Corpora

- Used in the documentation process for entry editing
- Used for automatic Bilingual Dictionary Drafting
- Used for the display of translated usage examples as part of the dictionary search result page (planned)

Literature Corpus

• Extracted from 81 German novels and their official translations to Basque, alignment hand-revisioned (Zubillaga & Sanz Villar 2011-2013, UPV-EHU), lemmatized and POS-annotated

Bible Corpus

• Aligned at verse level, lemmatized and POS-annotated

What can GFL learners expect from a bilingual dictionary?

- A bilingual dictionary should contain introducing and explaining prefaces in both languages. User interfaces of electronic dictionaries should be able to provide the possibility to switch between both languages as metalanguage, i.e. all instructions apart from the lemma signs, synonyms and translation equivalents should be available in both languages.
- German inflected forms should be searchable, information about the word form and a link to the corresponding canonical form listed as headword should be provided.
- Apart from single word lemma signs, also multi-word expressions such as light verb constructions or idiomatic phrases should be found in a dictionary, together with explaining instructions.
- A bilingual dictionary article should not only be furnished with insightful instructions for word sense (polysemy) disambiguation and a mapping to suitable Translation Equivalents, but also instructions related to morphology, syntax, and pragmatics. In the case of a dictionary for GFL learners these are the following:
 - Inflection morphology paradigm (or complete inflection tables) for the German verb, noun or adjective
 - For German verbs, auxiliary selection
 - Instructions related to valency or argument structure realisation
 - German synonyms
 - Frequency data

language pair. EURALEX, Bolzano.

- Pragmatics (Register)
- Collocates (planned)
- Phrases and Idioms (planned)
- Translated usage examples from bilingual corpora (in prep.)

References:

IDS (2009). Korpusbasierte Wortgrundformenliste DEREWO, v-40000g-2009-12-31-0.1, mit Benutzerdokumentation.

Lindemann, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. In Procedia - Social and Behavioral Sciences, 95, 249-257

Sciences, 95, 249–257.
Lindemann, D. (2014). Zweisprachige Lexikographie des Sprachenpaares Deutsch-Baskisch. In Domínguez Vázquez, M.J., Mollica, F. & Nied, M. (eds.)
Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik, Lexicographica Series Maior. De Gruyter, pp. 259–281.
Lindemann, D., Saralegi, X., San Vicente, I., Manterola, I. & Nazar, R. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density

Lindemann, D. & San Vicente, I. (in prep.). Euskarazko maiztasun lemategia gaurko teknologien ikuspuntutik. Bilbo: UPV-EHU.

De Schryver, G.-M. & Joffe, D. (2005). One database, many dictionaries-varying co(n)text with the dictionary application TshwaneLex. In Proceedings of the 4th

ASIALEX conference. Singapore, pp. 54–59.

De Schryver, G.-M., Joffe, D., Joffe, P. & Hillewaert, S. (2010). Do dictionary users really look up frequent words?—on the overestimation of the value of corpusbased lexicography. In Lexikos, 16.

www.ehu.es/eudelex

Microstructure

Following microstructures proposed in some dictionaries (e.g. Langenscheidt's bilinguals with German, PONS Großwörterbuch DaF), we organize the dictionary entry first and foremost syntactically, i.e. contrary to TEI Standard we define syntactical entities as parent, and sense groups as their child elements. For verbs, we discriminate (1) transitive, (2) intransitive, (3) reflexive or reciprocal and non-personal use as own syntactical entities.

Some arguments in favour of such an element order according to syntactic properties are:

1. It helps as a first orientation within longer dictionary entries

- 2. In text reception, it is a strategy of advanced German learners to identify the verb (the meaning of which they possibly don't know) and its arguments as syntactic entities and then to proceed to semantics
- 3. In German as a foreign language production, it is often not the meaning but the syntactic properties of a word dictionary users want to find information about.

aussetzen DerewoRank 1948 – trennbar

I Verb Transitiv +haben ►

1 (umea, animalia) utzietsi; (bazterrera) utzi; abandonatu

2 [+Dat.] eraginpean jarri

3 • eine Belohnung (auf etw.) <u>aussetzen</u> zbt.en saria iragarri

4 = unterbrechen eten

5 RECHTSWISSENSCHAFT atzeratu; geroratu

6 [an etw.] kezkatu; gaitzetsi

II Verb Intransitiv +haben >

1 eten • (eine Runde) <u>aussetzen</u> (txanda bat) galdu

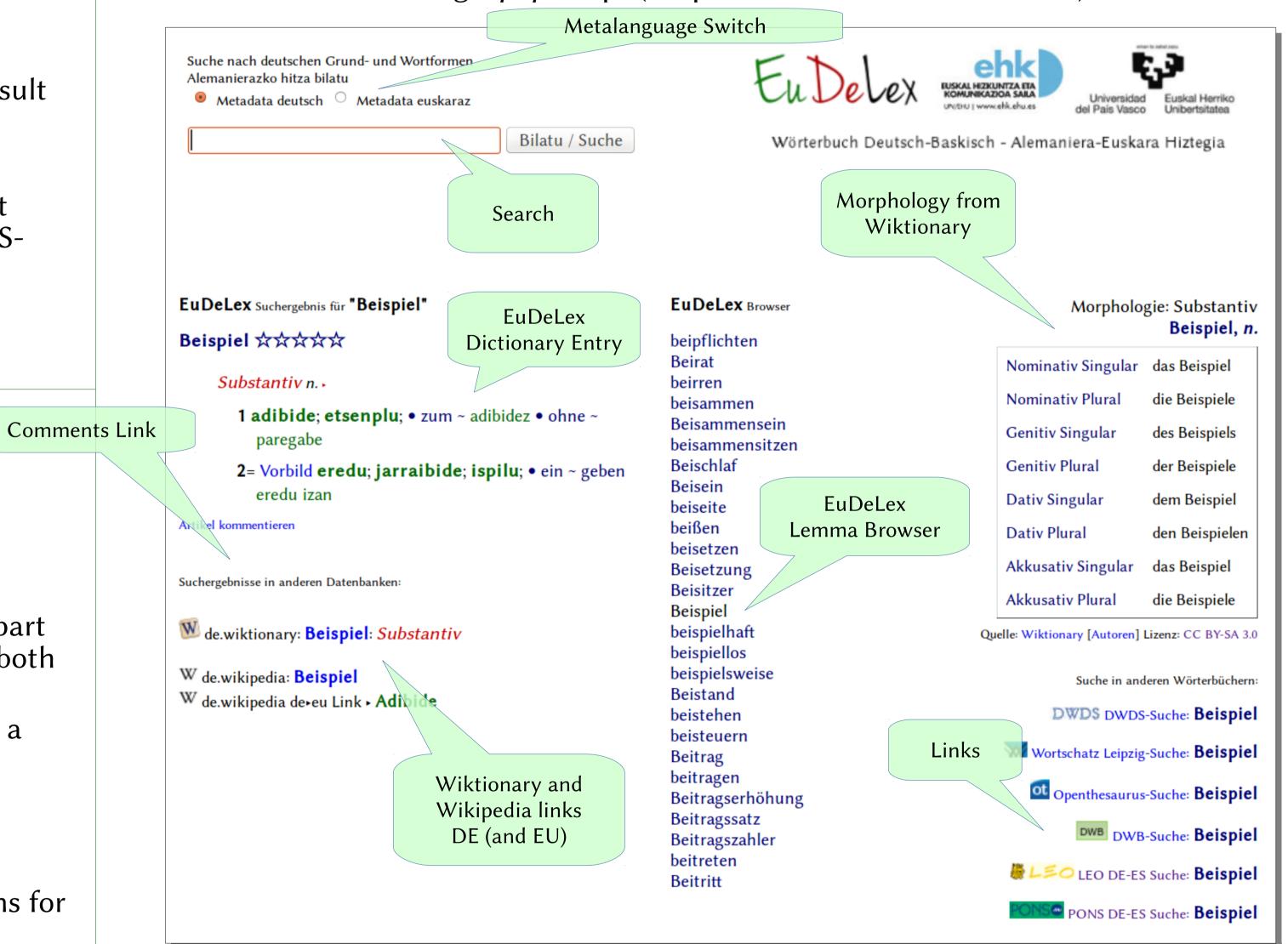
2 huts egin

Indications for German verb auxiliary selection are placed behind the syntactical flag. By valency formula, the user gets clues about a correct argument structure realization, which is helpful not only for text production, but also in text reception, in cases where a valency formula reveals the necessary to discriminate the German verb's polysemy.

Together with Basque translation equivalents, word sense groups are furnished with additional information such as specification of semantic domain and register, and with links to German synonyms. The structure of German noun, verb, adjective and adverb entries is presented in detail in Lindemann (2014).

Online Interface

TshwaneLex exports to XML format, the DTD of which is defined inside the application. Our own *perl* script transforms this XML to MySQL tables to be installed on the server; The User-Interface is based on a single *php* script (scripts: Lindemann & Nazar 2013).



In the page header, the user is prompted for a German headword to look for. The search result frame is organized in three columns:

- 1. Found dictionary entry or entries, and matching page titles, redirect pages and Basque translation links from the German editions of *Wiktionary* and *Wikipedia*.
- 2. Lemmabrowser: Links to direct neighbours on the lemma list;
- 3. Morphological information about the German headword, retrieved from Wiktionary: InflectionBelow, links to German monolingual and bilingual websites.

These three columns function independently, i.e. column (3) will also respond with search results if no entry is found in the local database. Inflected forms are also found in *Wiktionary*, the word form is then analysed, and a link to the canonical form is provided. As all data from *Wikimedia* is retrieved *on-the-fly* together with the local database query, it will always reflect the last version of these sources.



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