Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 249 – 257

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair

David Lindemann*

UPV-EHU University of the Basque Country, Tolosa Hiribidea, 70, 20018 Donostia, Spain

Abstract

Lexicography over the last decades has incorporated Corpus Linguistics methods. Lexicographers who start to work on an electronic dictionary, starting from scratch as Computational Linguists, and with little or no previous work done on their language pair, have to evaluate the contributions Corpus Linguistics methods may provide to their project, not only for lemmalist building, bilingual dictionary drafts and their documentation process in static entry editing, but also for dictionary publishing that contains dynamically generated corpus data displays. In the context of a low- or medium-density language pair, they will have to ask which electronic resources and tools are needed and available, and to evaluate the bilingual glossaries obtained with computational methods for their adequacy as bilingual dictionary draft. In this paper, we present some research on German-Basque corpus-based lexicography and describe our proposals for a new German-Basque electronic dictionary for Basque-L1 German learners. The observations and decisions to take at each stage of the lexicographical process may serve as reference for future dictionary projects that start from scratch, especially projects on other low- or medium-density language pairs.

Keywords: Lexicography; Bilingual Lexicography; Corpus Lexicography; Corpus Linguistics

1. Bilingual Lexicography and Corpus Linguistics

Lexicographers will find bibliographical reference that will introduce them into how to apply Corpus Linguistics methods to their work (*cf.* introductions in Atkins & Rundell, 2008: 53–96; Svensén, 2009: 43–58).

* Corresponding author. Tel.: +34-606-583-326; fax: +49-321-214-855-17.
E-mail address: david.lindemann@ehu.es

1877-0428 © 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CILC2013.
doi:10.1016/j.sbspro.2013.10.645

Today, nobody doubts that the so-called "corpus revolution" (e.g. in Rundell & Stock, 1992; Krishnamurthy, 2002; Hanks, 2012) has helped to make dictionaries reflect better the language in use. For example, it has been observed that pre-corpus dictionaries tended to take rare word senses (and translation equivalents) into account, "but missed important, common ones" (Klosa, 2007: 111). Now, frequency measurements are part of a standard lexicographical workflow, and word sketches that tell us about collocates and co-occurrences sorted by syntactical relations (Kilgarriff & Tugwell, 2002) can be instantly generated, provided that sufficiently large corpora and related NLP tools exist and are accessible for the language one is working on. But keyword concordances, frequency data and word sketches have not only become an indispensable resource to the lexicographers' documentation process in static dictionary entry editing: On more and more dictionary websites the static "editorial" dictionary entry is complemented with (or even replaced by) dynamically generated corpus-driven content.

1.1. Parallel Corpus Building and Density

Density, understood as "the availability of digitally stored material" in a language (Varga et al., 2005) is a factor not to be neglected in corpus-based lexicography. In most cases, the number of speakers of a language and its web size serve as approximation indicators for density, and the availability of electronic language resources is also a factor to have in mind. In the bilingual context, it is the density of the smaller partner in the language pair that determines by which methods parallel data can be gathered in order to set up an aligned translation corpus.

Table 1. Some density approximation indicators.

	German	Basque
Speakers	98 million	0,8 million
Biggest Corpus	5,4 billion	0,12 billion
Wikipedia Articles	1,6 million	0,15 million
ELRA Products	444	6

In our case, we have to deal with the fact that Basque, unlike German, is a low density language, that there is still not even a web domain (*.eus*) that corresponds to it, and that it is a minority language that is only co-official in just a part of the territory its speakers live in. As predicted by Varga et al. (*op. cit.*), it has therefore not been possible to gather parallel text by web mining, but there are still other parallel text sources, like literary text, religious text, movie dubs and subtitles or software localization files[1] that can be exploited for this purpose.

1.2. A new German-Basque Literary Parallel Corpus

At the University of the Basque Country, a German-Basque Literary Corpus has been created in the context of recent research for two PhD thesis in Descriptive Translation Studies (Sanz Villar, in press; Zubillaga, in press), using the content of 81 digital and/or digitalized and OCR-ed literary German originals and its official direct translations into Basque. In its actual version, the German-Basque translation corpus counts about 2 million tokens per language. The sentence alignment (146.000 sentence pairs) had to be done manually, starting from an automatic alignment at paragraph level, as more advanced sentence alignment tools capable of defining one-to-many-sentence pairings like *hunalign* need an initial bilingual glossary or "seed lexicon" to work with, in this case still not available.

The parallel corpus data has then been exported to TMX format and imported in the *SketchEngine* (Kilgarriff et al., 2004), where the German part could be lemmatized and POS-tagged with *TreeTagger*. In its first version, the Basque part of the parallel corpus had to be left in the stage of tokenization, which turned out to make the

[1] See http://de.glosbe.com/de/eu for a German-Basque corpus-driven dictionary based on software localization files.

computational tasks described below more difficult. Nevertheless, corpus queries in German by now can be done by lemma, which is the desired feature for dictionary entry editing and parallel KWIC display on the dictionary search result pages in the direction German to Basque.

1.3. Corpus-based Lemma List Retrieval

It has been shown 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). At the same time, frequency data is useful information for both dictionary editor and user. Therefore, it makes sense to build a lemma list starting from corpus-based frequency lists and to include frequency data in the published dictionary. For German, lemmatized frequency lists based on large reference corpora are available under public licenses (IDS, 2009). In the workflow for the new dictionary described in section 2, we compare the *DeReWo-40.000* list with human-revised lemmalists found in three editorial dictionaries, and delete, replace (adapt to a canonical form used as lemma in our macrostructure) or add German lemmata by hand. For Basque, the workflow is similar but differs in an important point: The Basque Language Academy *Euskaltzaindia* provides a 55.000 entry corpus-based lemma list under a public licence, each entry of which has been revised and approved by the Academy's lexicographical board. This basic lemmalist therefore needs no further editing, apart from possible complementation from other sources (for a survey of Basque corpora *cf.* Areta et al., 2008).

1.4. Corpus-based Semi-automatic Dictionary Drafting

Encouraged by research done on a medium-density language pair parallel corpus of a similar size (Hungarian and Lithuanian: Héja, 2010), we tried two word-alignment tools based on statistical methods, *GIZA++* (Och & Ney, 2003) and *Bifid* (Nazar, 2012). Our evaluation of the first-attempt results concludes as Héja (*op. cit.*: 2802) predicts: One the one hand, lemmatization of the corpus data is absolutely necessary for rich morphology languages like Basque, and on the other, a parallel corpus of 2 million tokens per language seems not to be enough for the training of the tested tools that rely on statistics and to obtain trustworthy translation equivalent pairings. Our forthcoming attempts will focus on building German-Basque *comparable corpora*, a task for which much more data can be gathered automatically from the web, and to extract bilingual glossaries by statistical methods from those. In a comparable corpus, if there is any initial alignment, it is done at document level. Texts that share similar topics will presumably share the same terminology, and therefore contain a significant amount of translation equivalents or interlinguistically alignable word senses. This can be true for news text from the same time period in different countries or wikipedia articles written in different languages describing the same subjects – a *parallel corpus* instead is defined as a collection of documents with aligned translations from a source to a target language. It is also planned to enrich the existing parallel corpus with more data exploiting the parallel text sources mentioned in section 1.1.

As stated before, a bilingual glossary is a useful and sometimes indispensable means in order to make some computational tools, such as advanced sentence and word aligners, work. Given the restrictions we still have to deal with in our case, until a bigger comparable or even parallel corpus is built and existing tools can be applied to it more efficiently, translation equivalent proposals or bilingual dictionary drafts can be obtained by other methods, such as extraction from *Wikipedia*'s interlanguage links (*cf.* Erdmann, 2007) and *Wiktionary*, which give results of bilingual glossaries with several thousands of entries.

The relatively best results so far we have obtained by aligning GermaNet/EuroWordNet (Hamp & Feldweg, 1997) and Basque WordNet/MCR (Gonzalez-Agirre et al., 2012) *synsets*: In the first attempt we aligned 16.000 Basque and 10.200 German Lexical Units to 8.500 parallel synsets or wordsenses. It has to be pointed out that the alignments obtained by this method link wordsenses, while all other methods described here don't include any word sense disambiguation.

Another current research (Saralegi et al., 2011, 2012) is carried out about building bilingual German-Basque glossaries by linking existing bilingual dictionaries through English as a pivot, the results of which are tested and then classified according to reliability by Inverse Consultation and corpus based methods (Distributional Similarity).

A detailed account of these different approaches and their results on German-Basque as language pair is planned for the near future.

1.5. Dictionary entry editing

In times of Corpus Lexicography, the lexicographers' introspection into themselves, that is to make use of their own linguistic competence, gets assisted by monolingual and bilingual corpus data displays. Although for some high-density language pairs there already exist examples of *corpus-driven bilingual dictionaries* that function without any human lexicographer being involved in the entry editing (but still don't disambiguate word senses)[2], in the production of *corpus-based bilingual dictionaries*[3] it is still human lexicographers who combine their own prejudice about how to explain a certain headword's polysemy to a dictionary user with their "reasoned condensations" (Grefenstette, 1998: 39) of corpus data displayed on their computer screen, distinguishing "information that is merely *true* from information that is *relevant*" (Rundell, 2002: 150) in order to code their conclusions into a dictionary entry. The discussion about the position of the human lexicographer's introspection, about what is left for it in times of Corpus Lexicography ranges from Sinclair's statement from 1985, according to which "evidence of secondary sources and the evidence of introspection should be brought in at a late stage in the process of compilation [...] the initial evidence should always be [...] from the observation of language in use" (cited in Krishnamurthy, 2008: 237f) on the one side, to the conclusion that despite large corpora and clever query tools being available, "the main resource remains the lexicographic knowledge of the project members in combination with large annotated text corpora which serve as a reference in all cases of doubt" (Introduction to GermaNet[4]), on the other.

As for the choice between hundreds or thousands of corpus citations that might be suitable usage and translation examples to be included in the dictionary entry, there already exist some assistant tools that rely on formal criteria (Kilgarriff et al., 2008), but also at this point it is still a human lexicographer who is more capable than a machine choosing and adapting a corpus citation to examples that apply to the needs of their dictionary users, especially in pedagogical lexicography (Kilgarriff et al., 2008: 431; Laufer, 1992).

1.6. Corpus concordances as part of dictionary search-result pages

Although first mentions of this possibility date from the 1990ies (Atkins, 1996: 523; Dickens & Salkie, 1996: 556), it is not until the very recent past that parallel corpus concordances have begun to show up on bilingual dictionary websites, and still there are not many examples.[5] Some monolingual dictionary websites already display KWIC lines and word sketches on their search result pages.[6] On the one hand, as mentioned above, it has been shown that corpora contain the common word senses and translation equivalents that in an editorial dictionary might get less attention than their frequency would suggest. On the other, human translators often choose unfrequent translation equivalents that are not listed in editorial dictionary entries. In Wolfgang Teubert's words, "the translator's design space is much larger than the language-neutral conceptual ontology (or the traditional bilingual dictionary) would leave us to believe" (Teubert, 2002: 203). Furthermore, not only lexically, but also syntactically, for the task of paraphrasing source language text bits in a way that will preserve meaning cross-linguistically and reflect the target language in use in an accurate way, human translators make use of different kinds of syntactical *transformations* (e.g. Nida & Taber, 2003), as can be appreciated in these citations from the German-Basque literary parallel corpus (see table 2 below).

---

[2] See, for example, http://dict.uni-leipzig.de

[3] We adopt here the distinction between *corpus-driven* and *corpus-based* given in Klosa (2007) to bilingual lexicography.

[4] See http://www.sfs.uni-tuebingen.de/lsd/germanet_structure.shtml

[5] See, for example, http://www.linguee.com

[6] See, for example, http://www.dwds.de/

Table 2. Parallel corpus citations for German headword *Ärger*.

German KWIC	Basque KWIC	comment
Aber auf dieser Route **hatte** ich noch nie **Ärger** mit ihnen!	Baina bide honetan ez **dut** inoiz haiekin **arazorik izan**!	V+N > V+N
Nun war der **Ärger** in Capricorns Stimme nicht mehr zu überhören	Dagoeneko nabaria zen Kaprikornioren ahotsean **haserrea**.	N > N
Meggie konnte nicht verhindern, dass ihre Stimme **vor Ärger** zitterte.	Meggiek ezin ekidin zezakeen ahotsak **amorruz** dar-dar egin ziezaion.	prep.+N > N+postp.
Meggie sah, wie sich Bastas Schultern **vor Ärger** spannten.	**Haserrearen haserrez**, Bastaren sorbaldak tenkatu egiten zirela ikusi zuen Meggiek.	prep.+N > N+postp. (idiomatic translation)
Mensch, biste nicht froh, daß du den ganzen **Ärger** los bist mit den Weibern?	Eta hi zer, ez al hago pozik, andreekin hituen **saltsa guztiak** bukatuta?	N > N (idiomatic translation)
Bastas Lippen wurden schmal **vor Ärger,** doch er verkniff sich eine Antwort...	Bastak ezpainak estutu zituen **haserre**, baina erantzuteko gogoari eutsi zion.	prep.+N > adv.

Parallel corpora of a significant size may indeed give us a general idea about context-sensitive and phrase-based human translation procedures that can hardly be predicted by a bilingual dictionary entry, and it is here where computational linguists look for "syntactically motivated transformation rules that explain human translation data" (Galley et al., 2004: 273). In an editorial dictionary entry, the headwords' features as syntactical entity, its part of speech, are usually maintained, if the target language allows it. In order to reflect the target language in use, it is therefore useful to provide real translation examples together with the usual bilingual dictionary entry as part of a search result. The display of parallel keyword concordances can be a suitable way to do so.

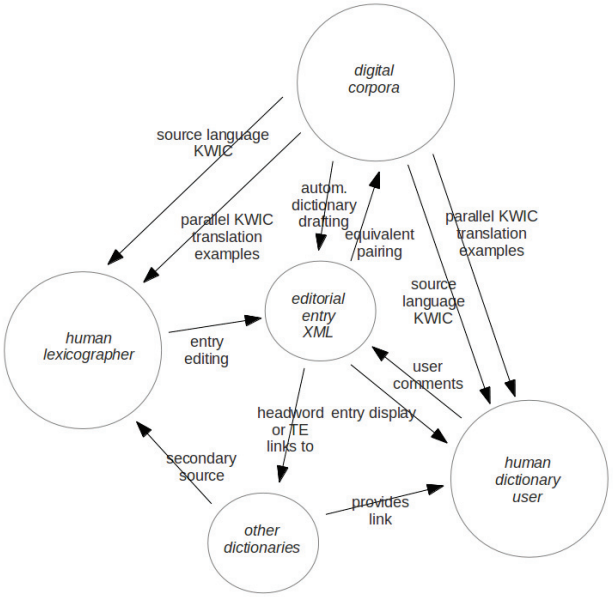

Fig. 1. Human lexicographer and digital corpora in bilingual lexicography

Figure 1 above shows the workflow links between a human lexicographer, digital corpora, other dictionaries as secondary source and a human dictionary user in electronic bilingual lexicography. Digital corpora provide to both lexicographer and dictionary user monolingual usage examples and translation examples. The human-edited

dictionary entry in our case has a feedback effect to the corpora and fills an important gap: A trustworthy bilingual glossary that may function as seed lexicon for corpus tools like sentence and word aligners can be extracted from these dictionary entries, so that the sentence and word alignment of the existing parallel corpus could be significantly enhanced. The described workflow is also retroactive at a second point: Dictionary users may improve the electronic dictionary entry by appending their comments to it.

2. A new German-Basque Electronic Dictionary

In the last decades, public and private education in the Basque Country has switched partly, and in some places entirely, to Basque as the language used in the lessons. For German as a foreign language, the first text book designed for Basque-L1 German learners written in German and Basque, without recurring to Spanish or French, was published in 2007 (Reuter & Wolff, 2007). In the following years, a series of studies about teaching German to Basque-L1 learners appeared (Braun, 2010; Reuter, 2010; Wolff, 2010). German teachers in the Basque Country have all made the same experience: Regardless of the teacher having recommended the use of monolingual dictionaries, learners stick to bilinguals, and lacking a suitable offer for German-Basque, they use Spanish-German dictionaries, which in terms of macro- and microstructure have a lot more to offer to them than the only available German-Basque pocket-size dictionary (Martínez Rubio, 2007). In order to fill this gap, it is the author's aim to propose and to provide an alternative, a German-Basque dictionary for Basque-L1 German learners[7]. We have started to work on the direction German-Basque, but don't discard the possibility to include a Basque-German part and to widen the scope of the dictionary and adopt its structure towards other functions than the described. Therefore, it makes sense to propose a database with an open structure that allows to include more data in the future, and to derive different dictionaries from it, a *Mutterlexikon* (Gouws, 2006) or *Multifunctional Lexical Database* as Pajzs understands it (Pajzs, 2009). For the design and editing of this database, we use the *TshwaneLex* Dictionary Writing System (De Schryver & Joffe, 2005). This application's XML output is transformed by a *perl* script into *html* and stored in a regular *MySQL* database.

2.1. Main components and features and compiling process

As for the macrostructure, we have begun to adapt the *DeReWo-40.000* list described in section 1.3 by comparing it to three human-made German dictionaries. In the first sample of 4500 lemmata, around 5% of the lemmalist differs from the initial list: 46 word forms have been adapted to a form used in our dictionary as lemma-sign. 217 have been deleted when missing in the secondary sources, most of which are large compound nouns and proper names. On the other hand, 186 lemmata that we regarded to be worth including have been added from the secondary sources.

For this sample of 4500 lemmata, the first in an alphabetical order ("A" - "beschleichen"), a first edition of bilingual entries is available now through the web interface. A combination of the lexicographers' own linguistic and lexicographic knowledge, the new German-Basque parallel corpus (see section 1.2) and several existing dictionaries has been the data source for entry editing. In *eudelex*, the German lemma is presented first as syntactical entity (see section 2.2 below); the polysemy is disambiguated on a second level, according to the Basque translation equivalents mapping to the German word senses, i.e. German word senses that share the same Basque translation equivalent are grouped together, although in some cases (of domain-specific terminology) the lexicographer decided to point out that a translation equivalent maps also in a specific domain, so that sense is listed as individual. 292 multiword expressions that appeared to be frequent according to the parallel corpus data, multiword expressions that include the lemma have also been listed in the entry, either below an existing word sense or as a word sense of its own.

In this first sample, usage examples and their translations are not included in the entries on a regular basis, although there have been added around 140 editorial (corpus-based or invented) examples in cases it seemed

[7] A preliminary version is available at http://www.ehu.es/eudelex

indispensable for a Basque-L1 German learner to understand the German lemma's polysemy. Until a lexicographer will have the chance to develop a standard criterion and edit suitable bilingual examples for all entries or at least for all verbs (which is a must in lexicography for learners), it is planned to perform on-the-fly queries to a corpus concordance tool and display parallel data from the existing German-Basque parallel corpus as parallel KWIC below the editorial entry.

As additional feature, we introduce in this preliminary version of the *eudelex* web interface the display of content from the German *Wiktionary* and *Wikipedia* presumably relevant to the Basque-L1 German learner, namely information about morphological features of the German lemma, apart from the links to the articles themselves. With every search performed by the user, *Wiktionary*'s API output is scanned for the morphological information contained in its corresponding article, on the one hand, and for Basque translation links, on the other. In the case of *Wikipedia*, it is the links to the German *Wikipedia* articles themselves, as well as any *interlanguage link* to a Basque *Wikipedia* article that the German page may contain.

Other microstructure elements found in this version of *eudelex* are not discussed here, such as German synonyms, cross-references and *valence formula* for German and Basque. A detailed presentation of *eudelex'* macro- and microstructure is in preparation (Lindemann, forthcoming).

2.2. Entry structure

A monolingual and bilingual dictionary entry structure often reflects the polysemy of the headword first and foremost. This implies to include the description of syntactical features of the headword like transitivity in "gram-groups" in each sense. According to TEI (Burnard & Sperberg-McQueen, 2007), a standard microstructure element hierarchy looks as follows; notice the information of syntactic properties below the semantic division as in this example from *Cambridge Online*:

Fig. 2. TEI semantics oriented vs. syntax oriented element hierarchy in *eudelex*.

In German learner's dictionaries (cf. for a survey Dentschewa, 2006) it is not uncommon to do it the other way round, and some authors explicitly propose to organize the hierarchy of microstructure elements as follows (Marello, 2010; Svensén, 2009: 278 citing Baunebjerg Hansen, 1990: 139): The polysemy appears as child elements of a syntactical entity such as "noun" or "adjective", "transitive verb", "intransitive verb" etc. The purpose of such an element order according to syntactic properties is to allow the dictionary user to focus on the part of the article that suits to the sentence they are trying to understand: In German text reception, language learners are taught to identify the verb and its arguments (the meaning of which they possibly don't know) as syntactic entities (subject, accusative object, dative object etc.) and then to proceed to semantics. In L2 production, it is often not the meaning but the syntactic properties of a word dictionary users want to be sure about: Can I use this verb as a transitive? Which auxiliary is selected by that verb in an intransitive sentence? As students' usual behaviour shows, they'd stick for that task to the bilingual dictionary they are familiar with, rather than to search for that kind of information in a monolingual dictionary, as their teacher might recommend.

3. Conclusions

Corpus methods have become a central issue for lexicographers. Corpora and corpus tools are not only indispensable in lemma list building, translation equivalent pairing and dictionary entry editing, but also an upcoming trend in dictionary publishing. When it comes to parallel corpus building, density is an important factor to bear in mind, as web-based automated corpus building methods only apply to language pairs with a considerable web size. One of the languages in our example, German, is definitely high-density, while Basque has to be regarded low or medium-density. Therefore, we apply other methods to build parallel corpora and to draft bilingual glossaries. These glossaries may serve for bilingual dictionary drafts as well as seed lexicon for sentence and word alignment tools.

Despite the growth of Corpus Linguistics and all kinds of corpora and related software, especially in the context of medium-density language pairs, human lexicographers' work is still far from being replaced by machines.

Acknowledgements

This study has been supported by the project IT665-13, funded by the Basque Government. Funding is gratefully acknowledged.

References

Areta, N., Gurrutxaga, A., & Leturia, I. (2008). Begiratu bat corpus-baliabideei. *Bat: Soziolinguistika aldizkaria*, *66*, 71–92.

Atkins, B. T. S. (1996). Bilingual dictionaries: Past, present and future. In M. Gellerstam, J. Jarborg, S.-G. Malmgren, L. Rogström, & C. R. Papmehl (Eds.), *Euralex'96 Proceedings* (pp. 515–546). Göteborg: Göteborg University.

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Braun, S. (2010). Die Rolle der Muttersprache im Deutschunterricht in zweisprachigen Gebieten am Beispiel des Baskenlandes. In C. Jarillot (ed.), *Bestandsaufnahme der Germanistik in Spanien: Kulturtransfer und methodologische Erneuerung* (pp. 25–36). Peter Lang.

Burnard, L., & Sperberg-McQueen, C. M. (2007). TEI P5: Guidelines for electronic text encoding and interchange. *TEI Text Encoding Initiative*.

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* (pp. 54–59). Singapore.

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 corpus-based lexicography. *Lexikos*, *16*(1).

Dentschewa, E. (2006). DaF-Wörterbücher im Vergleich: Ein Plädoyer für "Strukturformeln". In A. Dimova, V. Jesenšek, & P. Petkov (eds.), *Zweisprachige Lexikographie und Deutsch als Fremdsprache: drittes Internationales Kolloquium zur Lexikographie und Wörterbuchforschung, Konstantin Preslavski-Universität Schumen, 23. -24. Oktober 2005* (pp. 49–58). Hildesheim; New York: G. Olms.

Dickens, A., & Salkie, R. (1996). Comparing Bilingual Dictionaries with a Parallel Corpus. In M. Gellerstam, J. Jarborg, S.-G. Malmgren, L. Rogström, & C. R. Papmehl (Eds.), *Euralex'96 Proceedings* (pp. 551–559). Göteborg: Göteborg University.

Erdmann, M. (2007). *Extraction of Bilingual Terminology from the Link Structure of Wikipedia* (Master Thesis). Osaka.

Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004). What's in a translation rule. In *Proceedings of HLT/NAACL* (Vol. 4, pp. 273–280). Boston.

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual Central Repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul.

Gouws, R. H. (2006). Die zweisprachige Lexikographie Afrikaans-Deutsch - Eine metalexikographische Herausforderung. In A. Dimova, V. Jesenšek, & P. Petkov (eds.), *Zweisprachige Lexikographie und Deutsch als Fremdsprache: drittes Internationales Kolloquium zur Lexikographie und Wörterbuchforschung, Konstantin Preslavski-Universität Schumen, 23. -24. Oktober 2005* (pp. 49–58). Hildesheim; New York: G. Olms.

Grefenstette, G. (1998). The Future of Linguistics and Lexicographers: Will there be Lexicographers in the year 3000? In *Euralex '98 Proceedings* (pp. 25–41). Liège.

Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 9–15). Madrid: Association for Computational Linguistics.

Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography, 25*(4): 398–436.

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 the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valetta, Malta: European Language Resources Association (ELRA).

IDS. (2009). Korpusbasierte Wortgrundformenliste DEREWO, v-40000g-2009-12-31-0.1, mit Benutzerdokumentation. Institut für Deutsche Sprache, Programmbereich Korpuslinguistik.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX 2008*. Barcelona: Universitat Pompeu Fabra: 425-433

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX 2004* (pp. 105–116). Lorient, France.

Kilgarriff, A., & Tugwell, D. (2002). Sketching words. In M.-H. Corréard (Ed.), *Lexicography and natural language processing: a festschrift in honour of BTS Atkins* (pp. 125–137). Euralex.

Klosa, A. (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher. In W. Kallmeyer & G. Zifonun (Eds.), *Sprachkorpora. Datenmengen und Erkenntnisfortschritt* (pp. 105–122). Walter de Gruyter.

Krishnamurthy, R. (2002). The Corpus Revolution in EFL Dictionaries. *Kernerman Dictionary News*, *10*.

Krishnamurthy, R. (2008). Corpus-driven Lexicography. *International Journal of Lexicography*, *21*(3), 231–242.

Laufer, B. (1992). Corpus-based versus lexicographer examples in comprehension and production of new words. In *Proceedings of the Fifth Euralex International Congress* (pp. 4–9). Tampere: University of Tampere.

Lindemann, D. (forthcoming). *eudelex,* ein deutsch-baskisches elektronisches Wörterbuch. *Congreso Internacional Los diccionarios de alemán: evolución y nuevas perspectivas*,Valencia, October 2013

Marello, C. (2010). Verbos con construcciones tanto transitivas como intransitivas y/o pronominales en los diccionarios monolingües y bilingües italianos y españoles. In M. A. Castillo Carballo & J. M. García Platero (eds.), *La Lexicografía en su dimensión teórica* (pp. 411–434). Málaga: Universidad de Málaga.

Martínez Rubio, E. (2007). *Euskara alemana hiztegia*. Donostia: Elkar.

Nazar, R. (2012). Bifid: un alineador de corpus paralelo a nivel de documento, oración y vocabulario. *Linguamática*, *4*(2), 45–56.

Nida, E. A., & Taber, C. R. (2003). *The Theory and Practice of Translation* (4th ed.). Brill.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19–51.

Pajzs, J. (2009). On the Possibility of Creating Multifunctional Lexicographical Databases. In H. Bergenholtz, S. Nielsen, & S. Tarp (eds.), *Lexicography at a crossroads. Dictionaries and encyclopedias today, lexicographical tools tomorrow* (pp. 327–354). Bern: Lang.

Reuter, D. (2010). Interkulturelles Lernen am Beispiel baskischer Muttersprachler im Deutschunterricht. In C. Jarillot (Ed.), *Bestandsaufnahme der Germanistik in Spanien: Kulturtransfer und methodologische Erneuerung* (pp. 261–266). Peter Lang.

Reuter, D., & Wolff, J. (2007). *Deutsch - Euskaldunentzat*. Donostia: Erein.

Rundell, M. (2002). Good Old-fashioned Lexicography: Human Judgment and the Limits of Automation. In M.-H. Corréard (Ed.), *Lexicography and Natural Language Processing. A Festschrift in Honour of BTS Atkins* (pp. 138–155). Euralex.

Rundell, M., & Stock, P. (1992). The corpus revolution. *English Today*, *8*(04), 45–51.

Sanz Villar, Z. (in press). Hacia la creación de un corpus digitalizado, paralelo, trilingüe (alemán-español-euskera). In *Fraseología contrastiva del alemán y el español. Traducción y lexicografía* (pp. 43–58). München: Peniope.

Saralegi, X., Manterola, I., & San Vicente, I. (2011). Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 846–856). Association for Computational Linguistics.

Saralegi, X., Manterola, I., & San Vicente, I. (2012). Building a Basque-Chinese Dictionary by Using English as Pivot. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul.

Svensén, B. (2009). *A handbook of lexicography: the theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Teubert, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In B. Altenberg & S. Granger (Eds.), *Lexis in Contrast: Corpus-Based Approaches* (pp. 189–214). John Benjamins Publishing.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005* (pp. 590–596). Borovets.

Wolff, J. (2010). Sprachvergleich Baskisch/Deutsch und seine Auswirkungen für den Unterricht. In C. Jarillot (Ed.), *Bestandsaufnahme der Germanistik in Spanien: Kulturtransfer und methodologische Erneuerung* (pp. 37–42). Peter Lang.

Zubillaga, N. (in press). *Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez* (PhD Thesis). UPV-EHU, Vitoria-Gasteiz.