Beata Trawiński

Linguistic Data in Contrastive Studies

Addressing the Need for a Multilingual Parallel Resource Annotated with Semantic-Functional Information

Abstract: This paper argues for using authentic data not only as an empirical basis for linguistic generalizations but also for exemplification purposes in monolingual and particularly in bi- and multilingual contrastive studies. It shows that parallel data extracted from the available parallel corpora can – after enrichment with semantic-functional information while maintaining the available contextual, register-related and linguistic information – serve as a perfect data source for multilingual exemplification. Moreover, the analysis of semantic-functionally equivalent parallel sequences allows the investigation and exemplification of similarities and differences in how different languages express similar meaning from both a semasiological and an onomasiological perspective.

Keywords: parallel data, corpus, semantic-functional annotation, contrastive grammar, exemplification

1 Background: The Dual Function of Linguistic Data¹

Every linguistic work, whether a research paper or a grammar, whether written for scientific or didactic purposes, is inseparably tied to linguistic data. At the same time, linguistic data fulfill at least two different functions, both of which are subject to different restrictions and requirements. One function is to provide an empirical basis for linguistic generalizations, and the other is to exemplify these generalizations. Whereas the linguistic data used as an empirical basis is typically drawn from authentic language use (real performance), data serving as exemplification is usually constructed on the basis of such use.

There are good reasons for using constructed data for exemplification purposes: Constructed examples can be designed in a targeted way to have a clear focus on a specific problem and to be brief and concise, making it possible to save space, which is particularly relevant for print publications. However, constructed examples

¹ I would like to thank Piotr Bański, María José Domínguez Vázquez, Silvia Kutscher, Angelika Wöllstein and the audience of the international conference *Kontrastive Grammatikschreibung im europäischen Vergleich: Theorie, Methoden und Anwendungen*, held on 02.–03.10.2014 in Santiago de Compostela (Spain), for comments and discussion.

also have several drawbacks. Firstly, they are usually idealizations of real linguistic utterances and thus often appear artificial. Secondly, they are often introduced out of context and are unspecified with regard to register. Because of that, they lack explicit indications of appropriate usage and fulfill their epistemic function to a very limited extent.² In order to overcome these drawbacks and as a consequence of the empirical revolution in linguistics, more and more linguists tend to abandon constructed data in favor of authentic data. This trend can be observed in all areas of activity in linguistics, including research, teaching (see the widely discussed issue of *data-driven teaching / learning*), and grammar writing (see the grammar of English by Biber et al. (2002), which exclusively uses authentic texts and conversations as examples). Authentic language data are particularly advantageous in that they present real context- and register-related language usage. However, they are often not suitable for exemplification purposes because they are too complex and contain big portions of irrelevant material. The assets and drawbacks of authentic and constructed linguistic examples become even more obvious in contrastive research, in particular in the field of contrastive grammar writing, where more than one language is investigated and described and where the similarities and differences between the languages have to be adequately exemplified. It is examples that considerably determine the quality of a (contrastive) grammar (see also Weber 2005 and Rice 2007).

In this paper, we argue in favor of using authentic data³ for exemplification purposes and suggest how such data could be prepared to satisfy the needs of multilingual contrastive grammars. We show that parallel data extracted from the available parallel corpora can, after appropriate preparation, serve as a perfect data source for multilingual exemplification. In section 2 general characteristics of parallel corpora are described and the advantages and challenges related to the existing parallel corpora as well as the motivation for a resource annotated with semantic-functional information will be discussed. Section 3 presents the conceptual and practical issues concerning a multilingual parallel resource annotated with semantic-functional information intended to provide data for contrastive exemplification. Section 4 sums up the discussion and outlines future goals.

² See Stolz et al. (2011) for a discussion of the various functions of linguistic exemplification and the distinction between the descriptive and prescriptive functions of examples on the one hand and their epistemic function on the other hand.

³ By authentic data, we mean naturally occurring language data as opposed to made-up sentences. In this sense, translation data is a sort of authentic language data.

2 Parallel Data: Advantages and Challenges

For cross-linguistic research, including typology, contrastive linguistics and translation studies, bi- and multilingual electronic corpora are of particular relevance. Two types of bi- and multilingual corpora can be distinguished from the point of view of language selection: parallel and comparable corpora. Whereas parallel corpora consist of original texts in one language and their translations into another language (or languages) or texts which translate the same source,⁴ comparable corpora include exclusively original texts in two or more languages matched by a number of criteria such as topic or genre. In this paper, we focus on parallel corpora, arguing that they are particularly useful for cross-linguistic studies.

The major advantage of parallel corpora is that they provide – via translation equivalence – sequences of linguistic entities, such as words, word groups or sentences, in two or more languages that (i) convey the same meanings, (ii) appear in the same contexts, and (iii) are used in the same registers. It can thus be expected that the aligned sequences in two or more languages are functionally equivalent or at least approximate and thus qualify as an adequate *tertium comparationis* for a contrastive analysis (see James 1980; Chesterman 1998; Johansson 2007).⁵ Based on the functional equivalence of aligned sequences in parallel corpora, we can identify categorical and structural units in individual languages that are used to encode the relevant communicative functions. Sometimes we will be able to discover a one-to-one categorical and/or structural correspondence across languages. In other cases, we will find out that the same function is expressed by different formal means in different languages. Parallel corpora thus present a valuable linguistic resource for contrastive studies, as they allow us to gain insights into cross-linguistic similarities and differences that could be easily overlooked when working with monolingual corpora. This has been demonstrated in numerous contrastive papers, such as Altenberg (1999), Hasselgård (2007), Zufferey/Cartoni (2012) and Kaczmarek/Rosen (2013), to name just a few. There is also increasing interest in using parallel data in

⁴ In translation studies, these corpora are referred to as *translation corpora*.

⁵ Note that communicative-functional equivalence does not necessarily presuppose conceptual equivalence. Whether and to what extent functional equivalence involves conceptual equivalence is to be evaluated by linguists/grammar writers. However, examining functional equivalents in parallel texts is certainly one effective way to identify and reveal the conceptual dimensions underlying different languages.

typological research; see, for instance, van der Auwera et al. (2005), Da Milano (2007) and Wälchli (2007).⁶

At the same time, the number of bi- and multilingual parallel resources has been growing rapidly, and many of these resources are lemmatized and annotated morphosyntactically. The largest currently available multilingual parallel corpora include the Open Source Parallel Corpus (OPUS, <http://opus.lingfil.uu.se>) with 90 languages and 40 billion tokens (Tiedemann/Nygaard 2004; Tiedemann 2012), the parallel corpus InterCorp (<http://ucnk.ff.cuni.cz/intercorp/>)⁷ with 40 languages and 1.7 billion tokens (Čermák/Rosen 2012), the Acquis Communautaire (<https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis>) with 22 languages and over 1 billion tokens (Steinberger et al. 2006), and the European Parliament Proceedings Parallel Corpus (Europarl, <http://www.statmt.org/europarl/>) with 21 languages and 593 million tokens (Koehn 2005). Another noteworthy parallel corpus is the Parallel Corpus of Slavic and Other Languages (ParaSol, <http://www.slavist.de>) with 31 languages and 27 million tokens (Waldenfels 2006). Beyond that, there exists a variety of parallel corpora in a small number of languages, such as the Oslo Multilingual Corpus (OMC) with Norwegian, English, German, Dutch, Portuguese and French and the Stockholm MULTilingual Treebank (SMULTRON) with English, German and Swedish, as well as corpora in language pairs, such as the English-French corpus Hansards, the English-German translation corpus CroCo, the Polish and English Language Corpora for Research and Applications PELCRA, the Czech-English Parallel Corpus CzEng, the Slovene-English parallel corpus ISJ-ELAN, and many others.

All of the existing parallel corpora can already serve as a solid data source for contrastive research, but due to missing or sparse linguistic annotation they do not reach their full potential. This holds particularly for multilingual parallel corpora: Only two of them, namely InterCorp and ParaSol, are lemmatized and provide linguistic annotation. However, the linguistic annotation in InterCorp and ParaSol only covers morphosyntactic properties; functional-semantic annotation is not available.⁸ Moreover, linguistic information available in these corpora is language-

⁶ For a more general discussion on the use of parallel corpora in contrastive linguistics and language typology, see Cysouw/Wälchli (2007), Granger et al. (2003), Granger (2010) and Johansson (1999; 2007).

⁷ In fact, *InterCorp* consists of two parts: a manually aligned fiction part and fully automatically processed texts including political commentaries published by Project Syndicate and Presseurop, a package of legal texts of the European Union from the *Acquis Communautaire* corpus, the proceedings of the European Parliament dated 2007–2011 from the *Europarl* corpus, as well as movie subtitles from the *Open Subtitles* database.

⁸ There are a number of semantically annotated corpora. One such corpus is the TIGER/SALSA corpus, which is based on the syntactically annotated corpus of German newspaper texts TIGER and adds manual annotation of frame semantic roles as given by FrameNet (Erk et al. 2003). Another is

specific and thus not directly comparable across languages. As a result, only language-specific form-based queries can be performed and only research questions of the following kind can be modeled:

- What are the translation equivalents of lemma / word form / phrase W in language L1 in languages L2 ... Ln?
- Can lemma / word form / phrase W1 in language L1 be expressed by lemma / word form / phrase W2 in language L2?
- How can chains of the grammatical categories (or, more precisely category values) C1 ... Cn in language L1 be expressed in language L2?
- How can expressions bearing the grammatical features F1 ... Fn in language L1 be expressed in language L2?

Meaning- or function-related queries of the type below cannot be performed using the parallel corpora currently available.

- How is REFERENCE / QUANTIFICATION / REFLEXIVITY / POSSESSION / EXPERIENCE etc. expressed in languages L1 ... Ln?

However, precisely this kind of research question is substantial for cross-linguistic studies. Obtaining parallel sequences appropriately exemplifying specific functions across different languages from the available corpora using form-based queries is a laborious procedure. Moreover, the extracted data are usually more complex than is needed to illustrate the point at hand. For this reason, they are not suitable for unedited use as examples. With these facts in mind, we suggest building a multilingual database of parallel sequences annotated with semantic-functional information and designed especially for the purposes of data-driven contrastive research, in particular contrastive grammar writing, with a view to language didactic purposes. In doing so, we follow the guidelines for “good grammars” in Rice (2007, 152) with regard to examples:

“[...] Examples are of the utmost importance. There should be a wealth of data which is appropriate and authentic. The data should be carefully chosen to illustrate the point at hand. In addition, the data should be meticulously checked for spelling, and for consistency of spelling, glosses, and manner of glossing throughout the book.”

the TüBa-D/Z treebank, which has been manually annotated with senses from the German wordnet GermaNet (Henrich/Hinrichs 2013; 2014). However, with a few exceptions, semantically annotated corpora are monolingual.

In the next section, we sketch the conceptual and practical issues facing a resource that seeks to satisfy these criteria.

3 Towards a Multilingual Collection of Parallel Sequences Annotated with Semantic-Functional Information

We suggest to extract the data to be collected in the proposed resource out of the existing parallel corpora using the available corpus query tools. The (potentially randomized) extracted data should then be carefully examined by grammar writers and checked for quality of translation and accuracy of functional equivalence as well as for relevance and applicability as examples in a (bi- or multilingual) contrastive grammar. The appropriate data from this set should then be selected for inclusion in the new database. The selected data should include the largest possible portions of the extracted aligned data. This will facilitate retention of and access to the available contextual information. All of the meta-data information as well as the available linguistic annotation should also be retained in the new resource. Ideally, all extracted data would be linked to the source corpus. In the next step, the parts of the selected aligned data that are appropriate for exemplification purposes should be identified. Those parts should be additionally annotated with semantic-functional information. The new semantic-functional information should be linked to the relevant linguistic information adopted from the source corpora. To illustrate this idea, consider the following hypothetical research question in an English-German-Polish (functionally oriented) contrastive grammar: How is the concept of EMOTION (in the sense of an experiencer's emotions with respect to some content) linguistically realized and used in these three languages? The answer to this question is related to a number of semantic, structural and pragmatic issues, such as possible instantiations of the concept of EMOTION (e.g., *love*, *hate*, *dislike* etc.) or the number and kind of participants (the arity of the relation and semantic roles of the arguments). Furthermore, interrelations across different functional domains and

utterance levels might be considered, including morphosyntactic properties of the arguments and their syntactic realization (linking, syntactic function, omission, demotion etc.), internal temporal properties (Aktionsart, aspect), external temporal location (tense), the speaker's evaluation and judgment (mood), usage in particular register types etc.⁹ We approach our research question with a data-driven method by querying an English-German-Polish parallel corpus. In our example, the parallel corpus *InterCorp v8* was searched via the search interface *KonText* using the available query types, including CQL as well as lemma-, phrase- and word form-based queries. Recall that, due to the available annotation content, only form-based search is possible here. Figure 1 provides an excerpt from the query results obtained for the phrase-based query *in love* and the lemma-based queries *verlieben* and *zakochać* performed in the English, German and Polish subcorpus, respectively.¹⁰

The screenshot displays the KonText search interface with the following components:

- Header:** "kon text" logo, navigation links (Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, Help), and a corpus selector showing "corpus InterCorp v8 - English" with 139,641,916 positions.
- Search Summary:** Hits: 129 | 1 p.m. | 0.92 (related to the whole InterCorp v8 - English) | ARF: 46.63 | Result is shuffled | Selected lines: 3.
- Results Table:** A table with three columns for different corpora:
 - InterCorp v8 - English:** Contains search results for the phrase "in love". Examples include: "And I fell in love with him... we began having this love affair", "Well, I want to another universe and fell in love with a giant octopus and now I'm Pope of a new religion", "It has been quite rightly established, by Commission President Barroso amongst others, that people will not fall in love with the internal market... as important as it is, but want to see and enjoy cultural diversity, the cultural heritage of Europe", "I'm in love with a married pilot from Dallas", "I fell in love with you last winter", "Tish is in love", "Unless my ex-wife decides to fall in love with me again... or for the first time... yeah", "Anna Gulmiska... is not Jewish, but the 27-year-old brunette fell in love with Jewish culture... at school and university... and completed studies in Yiddish", "Or fall in love", "What drove her to love Goethe was not Goethe but the seductive image of the child Bettina in love with the old poet".
 - InterCorp v8 - German:** Contains search results for the lemma "verlieben". Examples include: "Und ich verliebte mich in ihn und wir hatten eine Liebesaffäre", "Nun, ich bin in ein anderes Universum und hab' mich in einen riesen Oktopus verliebt und nun bin ich der Papst einer neuen Religion", "Sehr zu Recht ist festgestellt worden, auch von Kommissionspräsident Barroso, dass die Menschen sich nicht in den Binnenmarkt verlieben werden... so notwendig er ist, sondern sie wollen kulturelle Vielfalt, den kulturellen Schatz Europas sehen und genießen", "Ich bin in einen verheirateten Piloten aus Dallas verliebt", "Ich habe mich im Winter in dich verliebt", "Ja, Tish ist verliebt", "Wenn meine Ex-Frau sich nicht wieder in mich verliebt... oder überhaupt ja... Nein!", "Doch sie hat sich... in der Schule und später denn am Ser Un... in die jüdische Kultur verliebt und hat Jiddisch studiert", "Oder dich in ihn verlieben", "Das... was sie in ihre Liebe zu Goethe hineintrieb... war nicht Goethe... sondern das verführerische Bild des Kindes Bettina... das in den alten Dichtern verliebt war".
 - InterCorp v8 - Polish:** Contains search results for the lemma "zakochać". Examples include: "A ja się zakochałam... i miałam romans", "Cóż... ucałował się do niego wziętych świata i zakochał się w wielkiej odmownicy... a teraz jestem papieżem nowej religii", "Całkiem słusznie ustalono, między innymi powiedział to przewodniczący Komisji... pan Barroso, że ludność nie zakocha się w rynku wewnętrznym... mimo, że ma on tak duże znaczenie... ale chcąc widzieć... korzystać z różnorodności kulturalnej Europy... i z jej dziedzictwa kulturalnego", "Zakochałam się w żonatym pilotce z Dallas", "Zakochałam się w tobie tej zimy", "Tish się zakochała", "Jestem miła była nie zakochałam się we mnie ponownie... albo raczej pierwszy raz... tak... Nie!", "Ale... w szkole... a potem na studiach... zakochałam się w kultuży żydowskiej, studiowałam jidysz", "Lubię zakochać się!", "Do miłowania Goethego nie skłonił jej Goethe... lecz czarujący obraz Bettiny dziewczę zakochanej w starym poecie".

Fig. 1: Excerpt from the query results for the queries *in love*, *verlieben* and *zakochać* in the English, German and Polish corpus in *InterCorp*.

⁹ See also Kutscher (2014) for a discussion of functional aspects of the linguistic realization of states of affairs in the spirit of Givón (1984).

¹⁰ We used the queries *in love*, *verlieben* and *zakochać* instead of *fall in love*, *sich verlieben* and *zakochać się* in order to guarantee more flexibility with respect to adjacency and word order.

The (randomized) corpus query results should then be inspected by grammar writers and the most representative of them should be selected for inclusion in the new database (see the selected line in figure 1). Portions of the selected parallel sequences that are directly relevant for answering and illustrating the research question (see the framed parts in the selected line in figure 1) should be reprocessed and enriched with semantic-functional information. The remaining parts of the selected sequences, which provide contextual background (which sometimes turns out to be rich and sometimes rather sparse), are still available, but can be hidden and may be accessed as required by the user.

The selected parallel sequences in figure 1 demonstrate a number of formal similarities and differences in the linguistic realizations of the concept EMOTION in English, German and Polish expressed by the relation FALL-IN-LOVE. For instance, the relation FALL-IN-LOVE is realized in English by a verbal multi-word expression (the light verb construction *fall in love*), while in German and Polish it is realized by a reflexive prefixed verb (*sich verlieben* and *zakochać się*, respectively). In all three languages, this relation is binary and includes the same types of participants: Experiencer and Stimulus. The Stimulus is realized by a prepositional phrase in English, German and Polish; the prepositional object is marked for accusative in German and locative in Polish. Furthermore, we can observe a number of phenomena at other linguistic, textual and functional levels, as well as at their interfaces. More precisely, in contrast to English and German, where the Experiencer is expressed by a first person singular pronoun, it remains syntactically unexpressed in Polish (*pro-drop*). Instead, in Polish the grammatical properties of the Experiencer (person, number and gender) are encoded morphologically on the predicate. Additionally, the Polish predicate is in the perfective aspect; there is no morphological aspect in English and German. Furthermore, the selected parallel sequences in figure 1 demonstrate different encodings of temporal event location. If we think of tenses as relations between time intervals S (Speech Time), R (Reference Time) and E (Event Time), following Reichenbach 1974), then the following interrelationships arise: $E=R \rightarrow S$ (past tense) in Polish and $E \rightarrow S=R$ (present perfect) in English and German. Whereas in Polish, this temporal relation is expressed by the morpheme *-ł-*, it is expressed analytically by the auxiliary verbs *haben* and *have* together with past participles in German and English.¹¹ Finally, the selected parallel sequences illustrate the usage of linguistic utterances incorporating the relation in question in fiction: The example is taken from (translations of) Stieg Larsson's novel *The Girl with the Dragon Tattoo*.

¹¹ Borsley and Rivero (1994) and Borsley (1999) analyze past tense forms in Polish as auxiliary constructions. However, Kupść and Tseng (2005) provide strong evidence against this approach. Here, we will not go into the details of this debate.

The example in figure 1 shows that a lot of features relevant for a contrastive analysis are located at different linguistic levels, such as morphology, syntactic structure and semantic-functional structure, and that individual properties are linked to each other across different levels. Of particular concern is the linkage between the form-related levels (morphology and syntax) and the meaning-related level (functional-semantic level). This kind of linkage is always intra-linguistic, which is due to the fact that all formal grammatical features are language-specific (see the notion of *descriptive categories* as opposed to *comparative concepts* discussed in Haspelmath 2007; 2010). On the other hand, there is an inter-linguistic linkage at the semantic-functional level resulting from the functional equivalence established in the process of translation. Figure 2 illustrates schematically the correlations between the form- and meaning-related levels across functionally equivalent expressions in different languages. As the schema indicates, the inclusion of language-specific formal features and cross-linguistic semantic-functional features in the representation of multilingual data as well as the linkage of the corresponding information across the levels and languages make it possible to look at the data both from a semasiological and onomasiological perspective.¹²

Fig. 2: Representation of and relationships between form- and function-based levels in multilingual parallel data

Given the complexity and heterogeneity of the information to be encoded in the new resource, data must be annotated at many levels, including sequence alignment, metadata, textual markup, and linguistic annotation. Linguistic annotation should

¹² For a discussion of a semasiological- and onomasiological-oriented approach in grammar writing, see Kutscher (this volume).

comprise the following layers, which can be adopted, if available, from the monolingual source corpora: word form, lemma, morphology, part of speech, constituent and dependency structure, and topological structure (for German). This part of the annotation is language-specific. In addition, language-independent, semantic-functional annotation, including predicate-argument structure, must be included.¹³ The semantic-functional annotation layer describes all the meaning components that underlie and motivate the alignment of particular sequences in different languages. Via language-independent semantic-functional information, language-specific form-related information can be accessed and linked across languages. At the language-particular level, the availability of both morphosyntactic and semantic-functional annotation layers makes it possible to map expressions and their meanings in two different directions: from a meaning/function to the formal/structural means by which it is expressed in a language (onomasiological perspective), and from an expression and its form/structure to its meaning/function (semasiological perspective).

Linguistic information in the new resource could be perfectly represented using typed feature structures, which would allow us to structure our data in a linguistically appropriate way and to establish linguistically-motivated identities between different substructures within expressions of a single language and across languages. The descriptive potential of feature structures is indicated by the (partial) description of a linguistic sign associated with the relation FALL-IN-LOVE in figure 3. Moreover, feature structures are a common data structure in many grammar formalisms, such as Lexical Functional Grammar, Head-driven Phrase Structure Grammar and Construction Grammar as well as grammar implementations, including parallel treebanks such as the ParGramBank (Sulger et al. 2013). It is thus possible to adopt solutions for encoding linguistic information from those frameworks and to use available language processing tools.

¹³ The annotation of the lexical meaning of predicates, their arguments and modifiers can, for instance, be based on the frame semantic paradigm (Fillmore 1985) and adopt frames and roles from the FrameNet database (Baker/Fillmore/Lowe 1998; Boas 2005; <<https://framenet.icsi.berkeley.edu>>). A number of projects have adopted (with modifications) the frames of the English FrameNet in resources for other languages, such as German (SALSA, Burchardt et al. 2006, <<http://www.coli.uni-saarland.de/projects/salsa>>), Spanish (Subirats/Petruck 2003), Japanese (Ohara et al. 2004, <<http://jfn.st.hc.keio.ac.jp/>>) or Polish (RAMKI, Derwojedowa et al. 2010, <<http://www.ramki.uw.edu.pl/en>>).

Fig. 3: Description of a feature structure modeling a sign associated with the relation FALL-IN-LOVE

The data in the resource can be encoded in XML according to the Guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative (TEI P5), with some modifications and extensions. For preliminary considerations regarding TEI encoding and the general architecture of the proposed resource, see Bański/Trawiński (2015).

4 Summary

In this paper, we discussed the dual function of linguistic data: as an empirical basis for (contrastive) linguistic generalizations on the one hand and as exemplification of linguistic generalizations on the other hand, and pointed out the assets and drawbacks of authentic and constructed linguistic data. Based on this discussion, we argue in favor of using authentic data not only as empirical evidence but also for exemplification purposes, particularly in bi- and multilingual contrastive grammars. Drawing on the assumption that the act of translation consists in the transfer of meaning from one language to another and that it results in functionally equivalent textual data in different languages, we suggest using authentic parallel data as examples in contrastive studies. At the same time, we point out the limitations of the existing multilingual parallel corpora, which only allow for language-specific form-based searches, and we argue for the need for a multilingual resource which makes cross-lingual contrastive browsing and data

querying possible. We propose to approach this goal via enriching the parallel data – carefully selected from the available corpora (along with existing metadata and available linguistic information) by grammar writers – with semantic-functional information. Adding semantic-functional information to parallel data would linguistically substantiate the alignment of the individual sequences. The analysis of semantic-functionally equivalent parallel sequences in the new resource would allow the investigation and exemplification of similarities and differences in how different languages express similar meaning from both a semasiological and an onomasiological perspective.

An implementation of the described resource idea (referred to as *CoMPaRS*, for *Collection of Multilingual Parallel Sequences*) is currently under conceptual and technical development within the project *German Grammar in European Comparison* (<http://www1.ids-mannheim.de/gra/projekte/gde-v.html>) at the Institute for the German Language in Mannheim.

5 Literature

- Altenberg, Bengt, *Adverbial connectors in English and Swedish: Semantic and lexical correspondences*, en: Hasselgård, Hilde/Oksemell, Signe (edd.), *Out of Corpora*, Amsterdam, Rodopi, 1999, 249–268.
- Baker, Collin F./Fillmore, Charles J./Lowe, John B., *The Berkeley FrameNet Project*, en: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – Volume 1*, Stroudsburg, Association for Computational Linguistics, 1998, 86–90.
- Bański, Piotr/Trawiński, Beata, *TEI encoding in the CoMPaRS project. Slides of the talk at LingSIG Meeting at TEI-MM in Lyon on 29th Oct. 2015*.
<http://sourceforge.net/projects/lingsig/files/Documents/SIG%20meeting%202015/LingSIG-2015_Banski%2BTrawinski.pdf/download> [último acceso: 12.06.2016].
- Biber, Douglas/Conrad, Susan/Leech, Geoffrey, *Longman Student Grammar of Spoken and Written English*, Harlow, Longman, 2002.
- Boas, Hans C., *Semantic Frames as Interlingual Representations for Multilingual Lexical Databases*, *International Journal of Lexicography* 18:4 (2005), 445–478.
- Borsley, Robert D. 1999. *Auxiliaries, Verbs and Complementizers in Polish*, en: Robert D. Borsley and Adam Przepiórkowski (edd.), *Slavic in Head-Driven Phrase Structure Grammar*, Stanford, CA: CSLI Publications, 29–59.
- Borsley, Robert D. and Rivero, Marfa Luisa. 1994. *Clitic Auxiliaries and Incorporation in Polish*. *Natural Language and Linguistic Theory* 12, 373–422.
- Burchardt, Aljoscha, et al., *The SALSA corpus: A German corpus resource for lexical semantics*, en: *Proceedings of LREC-06*, Genoa, 2006.
- Čermák, František/Rosen, Alexandr, *The case of InterCorp, a multilingual parallel corpus*, *International Journal of Corpus Linguistics* 17:3 (2012), 411–427.
- Chesterman, Andrew, *Contrastive Functional Analysis*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 1998.

- Cysouw, Michael/Wälchli, Bernhard, *Parallel texts: using translational equivalents in linguistic typology*, Sprachtypologie und Universalienforschung 60:2 (2007), 95–99.
- Da Milano, Frederica, *Demonstratives in parallel texts: A case study*, Sprachtypologie und Universalienforschung 60:2 (2007), 135–147.
- Erk, Katrin, et al., *Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation*, en: *Proceedings of ACL 2003*, Sapporo, Association for Computational Linguistics, 2003.
- Fillmore, Charles J., *Frames and the semantics of understanding*, Quaderni di Semantica 6 (1985), 222–254.
- Givón, Talmy, *Syntax. A functional-typological introduction*, Amsterdam/Philadelphia, J. Benjamins, 1984.
- Granger, Sylviane, et al. (edd.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Amsterdam/Atlanta, Rodopi, 2003.
- Granger, Sylviane, *Comparable and translation corpora in cross-linguistic research. Design, analysis and applications*, Journal of Shanghai Jiaotong University 2 (2010), 14–21.
- Haspelmath, Martin, *Pre-established Categories don't exist: Consequences for Language Description and Typology*, Linguistic Typology 11(2007), 119–132.
- Haspelmath, Martin, *Comparative concepts and descriptive categories in crosslinguistic studies*, Language 86:3 (2010), 663–687.
- Hasselgård, Hilde, *Using the ENPC and the ESPC as a parallel translation corpus: adverbs of frequency and usuality*, Nordic Journal of English Studies 6:1 (2007).
- Henrich, Verena/Hinrichs, Erhard, *Extending the TüBa-D/Z Treebank with GermaNet Sense Annotation*, in: Gurevych, Iryna/Biemann, Chris/Zesch, Torsten (edd.), *Language Processing and Knowledge in the Web*, Lecture Notes in Computer Science 8105, 2013, 89–96.
- Henrich, Verena/Hinrichs, Erhard, *Consistency of Manual Sense Annotation and Integration into the TüBa-D/Z Treebank*, en: Henrich, Verena/Hinrichs, Erhard/ de Kok, Daniël/Osenova Petya/Przepiórkowski, Adam (edd.), *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, Tübingen, University of Tübingen, 2014, 62–74.
- James, Carl, *Contrastive Analysis*, London, Longman, 1980.
- Johansson, Stig, *Corpora and contrastive studies*, en: Pietilä, Päivi/Salo, Olli-Pekka (edd.) *Multiple Languages – Multiple Perspectives*, AFinLA Yearbook 57(1999), 116–125.
- Johansson, Stig, *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*, Amsterdam/Philadelphia, John Benjamins, 2007.
- Kaczmarek, Elżbieta/Rosen, Alexandr, *Między znaczeniem leksykalnym a walencją – próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego* [Between lexical meaning and valency – towards a method for extracting equivalents based on a parallel corpus], Studia z Filologii Polskiej i Słowiańskiej 48 (2013), 103–121.
- Koehn, Philipp, *Europarl: A Parallel Corpus for Statistical Machine Translation*, en: *Conference Proceedings: the tenth Machine Translation Summit*, Phuket, Thailand, AAMT, 79–86.
- Kupść and Tseng 2005. *A New HPSG Approach to Polish Auxiliary Constructions* en: Stefan Müller (ed.), *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*. Stanford, CA: CSLI Publications, 253–273.
- Kutscher, Silvia, *Entwurf einer Makrostruktur zum Bereich der Sach-verhaltsversprachlichung im Rahmen des Projekts Grammatik des Deutschen im europäischen Vergleich (GDE-V)*, Arbeitspapiere der Abteilung Grammatik, Mannheim, Institut für Deutsche Sprache, 2014.
- Ohara, Kyoko, et al., *The Japanese FrameNet project: An introduction*, en: *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora at LREC 2004*, Lisbon.
- Reichenbach, Hans, *Elements of symbolic logic*, New York, Macmillan, 1974.

- Rice, Keren, *A typology of good grammars*, en: Payne, Thomas E./Weber, David J. (edd.), *Perspectives on Grammar Writing*, Amsterdam /Philadelphia, John Benjamins Publishing Company, 2007, 143–171.
- Steinberger, Ralf, et al., *The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages*, in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, 2006.
- Stolz, Thomas/Vossmann, Christina/Dewein, Barbara, *Kolonialzeitliche Sprachforschung und das Forschungsprogramm Koloniallinguistik: eine kurze Einführung*, en: Stolz, Thomas/Vossmann, Christina/Dewein, Barbara (edd.), *Kolonialzeitliche Sprachforschung. Die Beschreibung afrikanischer und ozeanischer Sprachen zur Zeit der deutschen Kolonialherrschaft*, vol. 1, Bremen, Akademie Verlag, 2011.
- Subirats, Carlos/Petruck, Miriam, *Surprise! Spanish FrameNet!*, en: *Proceedings of the Workshop on Frame Semantics at the XVII. International Congress of Linguists*, Prague, 2003.
- Sulger, Sebastian, et al., *ParGramBank: The ParGram Parallel Treebank*, en: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013, 550–560.
- Sykurska-Derwojed, Magdalena/Linde-Usiekiewicz, Jadwiga/ Zawislawska, Magdalena, *Projekt RAMKI: Rygorystyczna aplikacja metodologii kognitywno-interpretacyjnej (ram interpretacyjnych) do opisu polszczyzny*, en: Zawislawska, Magdalena (ed.), *RAMKI: Rygorystyczna aplikacja metodologii kognitywno-interpretacyjnej*, Elipsa, 2010, 7–16.
- Tiedemann, Jörg/Nygaard, Lars, *The OPUS corpus - parallel & free*, en: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, Lisbon, 2004.
- Tiedemann, Jörg, *Parallel Data, Tools and Interfaces in OPUS*, en: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, Istanbul, 2012.
- van der Auwera, Johan/Schalley, Ewa/Nuyts, Jan, *Epistemic possibility in a Slavonic parallel corpus – a pilot study*, en: Karlik, Petr/Hansen, Björn (edd.), *Modalität in slavischen Sprachen. Neue Perspektiven*, München, Sagner, 2005, 201–217.
- von Waldenfels, Ruprecht, *Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment*, en: Brehmer, Bernhard/Zdanova, Vladislava/Zimny, Rafal (edd.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9, München, 2006, 123–138.
- Wälchli, Bernhard, *Advantages and disadvantages of using parallel texts in typological investigations*, *Sprachtypologie und Universalienforschung* 60:2 (2007), 118–134.
- Weber, David J., *The linguistic example*, *Studies in Language* 30:2 (2005), 445–460.
- Zufferey, Sandrine/Cartoni, Bruno, *English and French causal connectives in contrast*, *Languages in Contrast* 12:2 (2012), 232–250.

- Rice, Keren, *A typology of good grammars*, en: Payne, Thomas E./Weber, David J. (edd.), *Perspectives on Grammar Writing*, Amsterdam / Philadelphia, John Benjamins Publishing Company, 2007, 143–171.
- Steinberger, Ralf, et al., *The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages*, in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, 2006.
- Stolz, Thomas/Vossmann, Christina/Dewein, Barbara, *Kolonialzeitliche Sprachforschung und das Forschungsprogramm Koloniallinguistik: eine kurze Einführung*, en: Stolz, Thomas/Vossmann, Christina/Dewein, Barbara (edd.), *Kolonialzeitliche Sprachforschung. Die Beschreibung afrikanischer und ozeanischer Sprachen zur Zeit der deutschen Kolonialherrschaft*, vol. 1, Bremen, Akademie Verlag, 2011.
- Subirats, Carlos/Petruck, Miriam, *Surprise! Spanish FrameNet!*, en: *Proceedings of the Workshop on Frame Semantics at the XVII. International Congress of Linguists*, Prague, 2003.
- Sulger, Sebastian, et al., *ParGramBank: The ParGram Parallel Treebank*, en: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013, 550–560.
- Sykurska-Derwojed, Magdalena/Linde-Usiekiewicz, Jadwiga/ Zawislawska, Magdalena, *Projekt RAMKI: Rygorystyczna aplikacja metodologii kognitywno-interpretacyjnej (ram interpretacyjnych) do opisu polszczyzny*, en: Zawislawska, Magdalena (ed.), *RAMKI: Rygorystyczna aplikacja metodologii kognitywno-interpretacyjnej*, Elipsa, 2010, 7–16.
- Tiedemann, Jörg/Nygaard, Lars, *The OPUS corpus - parallel & free*, en: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, Lisbon, 2004.
- Tiedemann, Jörg, *Parallel Data, Tools and Interfaces in OPUS*, en: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, Istanbul, 2012.
- van der Auwera, Johan/Schalley, Ewa/Nuyts, Jan, *Epistemic possibility in a Slavonic parallel corpus – a pilot study*, en: Karlik, Petr/Hansen, Björn (edd.), *Modalität in slavischen Sprachen. Neue Perspektiven*, München, Sagner, 2005, 201–217.
- von Waldenfels, Ruprecht, *Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment*, en: Brehmer, Bernhard/Zdanova, Vladislava/Zimny, Rafał (edd.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9, München, 2006, 123–138.
- Wälchli, Bernhard, *Advantages and disadvantages of using parallel texts in typological investigations*, *Sprachtypologie und Universalienforschung* 60:2 (2007), 118–134.
- Weber, David J., *The linguistic example*, *Studies in Language* 30:2 (2005), 445–460.
- Zufferey, Sandrine/Carioni, Bruno, *English and French causal connectives in contrast*, *Languages in Contrast* 12:2 (2012), 232–250.