Innovations in parallel corpus search tools

Volk, Martin ; Graën, Johannes ; Callegaro, Elena

Abstract: Recent years have seen an increased interest in and availability of parallel corpora. Large corpora from international organizations (e.g. European Union, United Nations, European Patent Office), or from multilingual Internet sites (e.g. OpenSubtitles) are now easily available and are used for statistical machine translation but also for online search by different user groups. This paper gives an overview of different usages and different types of search systems. In the past, parallel corpus search systems were based on sentence-aligned corpora. We argue that automatic word alignment allows for major innovations in searching parallel corpora. Some online query systems already employ word alignment for sorting translation variants, but none supports the full query functionality that has been developed for parallel treebanks. We propose to develop such a system for efficiently searching large parallel corpora with a powerful query language.

Innovations in Parallel Corpus Search Tools

Martin Volk, Johannes Graën, Elena Callegaro

University of Zurich
Institute of Computational Linguistics & English Department
volk@cl.uzh.ch, graen@cl.uzh.ch, elena.callegaro@es.uzh.ch

Abstract

Recent years have seen an increased interest in and availability of parallel corpora. Large corpora from international organizations (e.g. European Union, United Nations, European Patent Office), or from multilingual Internet sites (e.g. OpenSubtitles) are now easily available and are used for statistical machine translation but also for online search by different user groups. This paper gives an overview of different usages and different types of search systems. In the past, parallel corpus search systems were based on sentence-aligned corpora. We argue that automatic word alignment allows for major innovations in searching parallel corpora. Some online query systems already employ word alignment for sorting translation variants, but none supports the full query functionality that has been developed for parallel treebanks. We propose to develop such a system for efficiently searching large parallel corpora with a powerful query language.

Keywords: Parallel Corpora, Query Languages, Word Alignment

1. Introduction

Translated documents in multiple languages (here: parallel documents) are highly regarded as valuable resources for various tasks in natural language processing and linguistic research. Parallel corpora are useful for fields as diverse as word sense disambiguation, terminology extraction and contrastive corpus linguistics. The usefulness of these resources for contrastive linguistics, in particular, has increased tremendously with the possibility to automatically align the texts not only on the sentence level, but also on the word level. Statistical word alignment (Och and Ney, 2003, Tiedemann, 2011) has made it possible to align large text collections efficiently.

Now the challenge lies in accessing this wealth of linguistic information in the best possible way for various user groups. Language learners want to get typical translations and usage contexts. Similarly, translators are interested in translation variants with their frequencies and usage in different genres or domains. Linguists, in addition, want to access special phenomena that emerge from language contrast, and they might be interested in diachronic developments.

This paper presents a comparison of various online search tools for parallel corpora. We argue that automatic word alignment enables new search options that are interesting for translators and linguists alike. Some search tools are commercial and well-established, others are experimental and provide insights in new application scenarios and visualization options.

2. User Groups and Usage Examples

For this study, we distinguish three types of parallel corpus users. First, there are language learners who want to get translation suggestions for a certain source language word or phrase. Traditionally, this group has consulted bilingual dictionaries, printed or electronic, which contain manually structured information on the source and target language sides. For example, the dictionary entry may contain part-of-speech information, the inflectional paradigm, translation options and few selected usage examples. Parallel corpora now offer a virtually limitless array of usage examples and - if word-aligned - also allow for prominence rankings based on frequency counts. For instance, when searching for the German word *Handy* in a collection of 1 million contemporary German - English TV subtitles, we find that *cellphone* and *phone* are equally frequent whereas *mobile phone* is less often used. We also find evidence that the abbreviations *cell* and *mobile* are also popular options in colloquial speech.

Admittedly, this usage of parallel corpora requires a basic level of understanding in the target language. Beginners in a new language might still be served better by a dictionary. But for the medium to advanced language learner or second-language user, the advantages of parallel corpora are apparent.

The needs for professional translators, our second user type, are similar. They want to look up words quickly, understand usage contexts and preferences, and employ this knowledge for new translations. For some time, translation workbenches have provided parallel concordances as views into the translation memory. But these concordances were only presented as parallel sentences which leaves the translator who is searching for a specific source language term with the task of identifying the corresponding term in the target language sentence. In contrast to language learners, translators are much more interested in finding genre- and domain-specific translation variants. It is therefore of central importance that parallel corpus search systems for translators allow for genre and domain filters.

The third user group, linguists (including scholars in translation studies), are the most demanding users of corpus query systems. They would like to formulate precise queries with reference to linguistic annotations and alignments. For example, a linguist who studies zero article use in English might want to search for German noun phrases with a determiner but only if the corresponding English noun phrase does not contain a determiner. Currently, such detailed requests can only be processed by specific query systems, as e.g. the query system over parallel treebanks described in section 4.5. below. But in treebanks the amount

of parallel text is limited, and none of the online search systems over large parallel corpora supports such queries.

3. Related Work

Our work is related to others in various dimensions. One related dimension is monolingual corpus search tools. There are a number of powerful search tools for monolingual corpora, the IMS Open Corpus Workbench[1] is probably the best known. It was developed in the 1990s and is used by many researchers around the world. The Workbench contains a flexible query processor CQP that accepts regular expressions and complex search queries. It allows to manage large corpora and to query them efficiently. Unfortunately, it does not support the management of parallel corpora.

Another well-known monolingual tool with a powerful query language, called CQL, is the XAIRA system for the XML-encoded version of the British National Corpus. Similar to CQP, CQL allows for logical combinations of search patterns which have access to all annotated attributes on word, phrase and sentence levels. The application of regular expressions to said patterns in combination with complex boolean expressions provides the user with a powerful means for retrieving matches not only from the BNC but from any kind of XML-encoded (corpus) data.

Some of the early work on using parallel corpora for contrastive linguistics originates in Norway. It started with the English-Norwegian Parallel Corpus, further languages (German, Dutch and Portuguese) were added at a later stage, and the extended version of the corpus is now called the Oslo Multilingual Corpus (OMC). Specific tools like the Translation Corpus Aligner (Hofland and Johansson, 1998) and the Translation Corpus Explorer (Ebeling, 1998) were developed for this project. The former focused on language-specific information, and used the technique of so-called anchor words (or anchor list); the criteria taken into consideration to build this anchor list were the frequency of the words and their equivalence in the two languages. The Translation Corpus Explorer "operates on a database, which was generated by the texts. There is a separate program which loads the texts into a database, and this program must be run before any searching can take place" (Ebeling, 1998) p.104. The interface and the basic search facilities enable lexical searches; the browser shows then the first sentence where this word is used and the corresponding sentence in the other language(s), in parallel texts. Around that time, (Lawson, 2001) argued for the role of parallel corpora in language learning. She quotes (Teubert, 1996) as saying "a parallel corpus of a reasonable size contains more knowledge about translational equivalents than any bilingual desk dictionary". She continues (p. 286): "It is to be expected and hoped that the current generation of electronic bilingual dictionaries largely based on traditional print dictionaries will be superseded by resources allowing access on demand to material drawing on multilingual corpora." This is exactly what has happened in recent years.

When Lawson wrote her paper in 2001, ParaConc, a PC-based system, was the most prominent search tool for parallel corpora (Barlow, 2002). ParaConc allowed users to load parallel text files, to align them on the sentence level and to do corpus linguistic searches (like KWIC, collocations etc.) on one side of the corpus which resulted in displays of the search language sentences with their aligned translated sentences. ParaConc did not support word-alignments, but it was popular throughout a decade and, for instance, still used by (Cartoni and Lefer, 2011) to extract trilingual concordances.

Among the few other tools to search parallel corpora since then, most of them targeted at translators. Translator workbenches by companies like SDL/Trados, Across, and STAR allow searching through translation memories. However, when the user queries such a database of parallel sentences, she has to locate the corresponding word in the translated sentence herself. This is cumbersome and, moreover, it does not allow to sort the translation examples according to their variants. Only a word-aligned parallel corpus tool enables us to efficiently retrieve desired information of this kind.

Such search tools have become available in recent years. (Bourdaillet et al., 2010) investigated the issue of using automatic word alignment for a commercial bilingual concordancer. They discuss various (statistical) methods for filtering bad alignments (e.g. a noun aligned with a determiner) and merging translation variants (e.g. singular and plural nouns or different verb forms). The paper includes a detailed evaluation using 2000 queries on the 8.3 million English-French sentence pairs from the Canadian Hansard. Web services that allow similar functionality are Glosbe, Linguee and Tradooit, all meant as online dictionaries with usage examples and not as tools for linguistic research. In order to experiment with online searches over genre-specific parallel corpora we have developed Bilingwis (see section 4.4.), a tool that currently enables the searching of different parallel corpora of mountaineering texts, law texts, and TED talks. One of its special features is support for searching German separable prefix verbs which are discontiguous in the sentence.

In previous papers we have argued that word-aligned parallel corpora provide new insights into parallel texts (Volk et al., 2011a, Bywood et al., 2013). All kinds of contrastive language studies have profited from parallel corpora (Borin, 2002, Johansson, 2007, Oakes and Ji, 2012, Aijmer and Altenberg, 2013), with great care taken to distinguish between original texts and translated texts. In some of the large multilingual corpora (like e.g. Europarl) the translation direction cannot always be retraced. This limits a corpus' usefulness for linguistic studies.

In most contrastive studies, the parallel corpus was accessed through locally developed special-purpose programs (as e.g. the "Perl Translation Corpus Explorer"[2] for searching the Oslo Multilingual Corpus). We argue that web-based, general-purpose parallel corpus query systems will facilitate access and allow for a much broader user base.

4. Parallel Concordancing Systems

In this section we sketch the criteria for comparison of parallel concordancing systems and subsequently list some re-

[1] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

[2] see http://www.hf.uio.no/ilos/english/services/omc/UsingTCE.htm

cently developed systems. When introducing the systems we will comment on the following aspects.

1. Number of language pairs, the size of the parallel corpora (usually given as the number of tokens per language), and the type of the parallel corpora (e.g. genre and domain, professionally translated vs. crowd translated, pre-aligned via subtitle time codes or translation memories);

2. Type of metadata (e.g. original language vs. translation; text creation date);

3. Type of linguistic annotation (Part of speech tagging, lemmatization, recognition of multiword expressions (e.g. named entities, verb particle pairs, support verb units), chunks or parse structures);

4. Expressiveness of the Query Language (wildcard search, regular expression search, multiword search, cross-language search, negated search) and search efficiency;

5. Ease of user interface, user feedback options (e.g. for the disapproval of incorrect alignments), visualization of search results (e.g. KWIC views, context views, highlighting), system suggestions (e.g. "see-also" links to similar words or expressions).

4.1. Tools for Language Learners: Glosbe, Linguee, Tradooit

Glosbe[3] calls itself a "multilingual online dictionary". It supports a large number of languages, for example it offers 125 languages in combination with German, from Afrikaans-German to Zulu-German. For languages with parallel corpora Glosbe also shows word-aligned usage examples. Glosbe focuses more on "recall" than precision and thus often highlights incorrect or mis-aligned words. For example, when searching from Spanish to German with the term "zumo de ova" (EN: grape juice), the system highlights any occurrence of any token from the search term in the source language. On the positive side, each usage example comes with a symbol indicating its origin, but the user interface does not allow sorting according to translation variant or source text.

Similar but more carefully crafted systems are Linguee and Tradooit. Linguee[4] is a commercial system, online since 2009, and supports searches in any bilingual combination of English, French, German, Portuguese and Spanish plus combinations of Chinese, Japanese, Russian and others with English. Linguee works on word-aligned texts and provides two views, one in combination with dictionary information and one where hits are sorted according to target language words. In addition to example sentences, Linguee presents manually prepared dictionary information (including the parts of speech for source and target words).

Tradooit[5] is a Canadian system that supports searches in bilingual texts in English - French (370 million words) and in English - Spanish (260 million words). The site gives a nice overview of the corpora in Tradooit with their respective sizes in terms of segments and words. The translation variants are sorted according to frequency and nicely highlighted in the example sentences. The user can filter the results according to sources (Europarl, EU bookshop, OpenSubtitles, Canadian Hansard, Statistics and Environment sites by the Government of Canada). The user can also view each found sentence pair in the context of the complete sentence-aligned original bitext.

In both Linguee and Tradooit, the user can mark his approval or disapproval of the alignment by clicking a thumb-up or thumb-down button. This is a welcome functionality since many automatic alignments on the sentence level and subsequently on the word level are incorrect.

All three systems allow for multiword queries but are imprecise in highlighting the corresponding target words then. We tested the systems for English to German (Tradooit for English - Spanish) with three verb-particle pairs *fool around, knock down, speed up* and three noun compounds *oil tanker, lung cancer, board meeting*.

For example, for *fool around* Glosbe delivers 4 German translations (*Quatsch machen, Unsinn machen, blödeln, herumalbern*), 4 English explanations, an automatic translation from Google Translate, and 4 suggestions for "similar terms" (*fooled around, fooling around, stop fooling around, to fool around*). The subsequent usage examples in Glosbe hardly ever highlight the correct term on the German side. The other verb-particle pairs result in even more translation variants from various dictionary sources, but also lack the correct highlighting in the examples. Glosbe knows all three noun compounds, but strangely enough in the case of *board meeting* it delivers usage examples where these two words are not adjacent in English which renders these examples useless.

Linguee also outputs 4 German translations for *fool around* (= *Quatsch machen, Unsinn machen, albern, herumalbern*). In addition, it offers translations for *fool* and *around* in separation. Linguee presents usage examples but with mostly erroneous highlighting. For *oil tanker* the Linguee dictionary tells us that it should be translated into German with *Öltanker*, but then the first two usage examples show examples of *Öltankschiff* which is a good alternative. Linguee makes up for this by listing synonyms for all parts of the compound where we find that *tanker* can also be translated as *Tankschiff*.

For *oil tanker* Tradooit presents the Spanish translations *petrolero* and *naufragio de petrolero*, but it also presents the incorrect *accidente de petrolero* (EN: *oil tanker accident*) twice. Both hits show the same source - target sentence pair. Obviously, the Tradooit team has not been careful in removing duplicates.

In summary, Glosbe has by far the largest coverage in terms of language pairs, followed by Linguee. Both Glosbe and Linguee present dictionary information when available. Tradooit has an advantage when genre filtering is wanted by translators.

[3] see http://www.glosbe.com

[4] see http://www.linguee.com

[5] see http://www.tradooit.com

4.2. Translator Search Tools: TAUS Search

There is one other online search system that is clearly geared towards translators. TAUS[6] is meant for terminology searches in domain-specific translations. It is based on Translation Memory data provided by translation agencies. All texts are word-aligned. TAUS supports more than 50 languages (plus varieties as e.g. 4 types of English and 3 types of Spanish). It stores large parallel collections, e.g. the language pair English (US + UK) to French (France) has 316 million words in a direct translation memory and another 180 million words in matrix translation memory that TAUS has created via a pivot language.

TAUS Search comes with a number of interesting settings. For example, TAUS allows search restrictions for industry (7 options: Computer Software, Legal Services, Pharmaceuticals and Biotechnology, ...), and for genre (content type; 6 options: Instructions of use, Sales and marketing material, Standards, statutes and regulations, ...). In addition, it allows (simple) Part of speech (adjective, noun, verb) and lemma searches. The accuracy of the PoS search is obviously limited by the precision of the PoS tagger. We checked for *can* as a noun and *house* as a verb. The result for *can* was disappointing. Out of the 20 hits there were 3 cases where it functions as a name (like in *CAN messages*), all others were verb occurrences (some of them admittedly with spelling errors in the neighboring words). The search results for *house* as a verb, on the other hand, are very good, 19 correct hits, and clearly demonstrate the usefulness of PoS tagging.

As a special feature, TAUS provides information on the direction of the human translation for each sentence pair. Then users know whether they are looking at authored or translated text.

4.3. Linguistic Search Tools: ParaSol, ParaQuery

Linguistic search tools allow queries over linguistic annotation levels such as PoS tags or morphology labels. ParaSol[7] is such a search tool through a large collection of literary texts in a variety of slavic, germanic, romance and some other languages (von Waldenfels, 2011). As a special feature it delivers hits in multiple languages in parallel which enables the user to compare translations across three or more languages in a single view.

Paraquery[8] allows searches in English-German parallel corpora (Europarl and United Nations) and presents search results with keywords in context on the source side plus aligned sentences on the target side. Paraquery supports searches over PoS tags on the query language side. Neither of these two systems includes word alignment.

4.4. Experimental Systems: OPUS Word Search, Bilingwis

OPUS is a large collection of freely available parallel corpora by (Tiedemann, 2012). It features EU and UN corpora such as Europarl, JRC, European Central Bank and MultiUN, as well as subtitles (OpenSubtitles and TED) and technical documentation (KDE, OpenOffice, PHP). Part of this material can be searched online in more than 20 languages[9].

The OPUS Corpus Query is a multilingual concordance system based on the Corpus Query Workbench. In the standard setting this OPUS system presents hits in the source language plus in any number of parallel corpora, however without word alignment. A second system called OPUS WordAlign compensates for this and allows searches in word-aligned versions of the EU constitution, Europarl and OpenSubtitles. Unlike most other systems mentioned above OPUS WordAlign is able to list translation variants in two or more languages simultaneously.

Our group has built Bilingwis (bilingual word information system)[10], a web-based search tool for word-aligned parallel corpora (Volk et al., 2011a). Bilingwis allows searches for word forms or for lemmas as many of the other systems. At the moment, Bilingwis is available for the language pairs German-French (over alpine texts and the Swiss federal laws), English-Chinese (over TED talks) and German - Rumansh (Swiss laws and press texts) (Weibel, 2014).

Bilingwis is a domain-specific search tool. For example, its documents for the language pair German-French are from two domains. We have built one system on 5 million words of mountaineering texts from the Swiss Alpine Club and another one on the collection of Swiss law texts of about the same size which allows for interesting contrasts. Domain-specific systems enable the user to search for specific terms such as German *Steigeisen* (EN: crampon) or French *sommet* (EN: summit).

Bilingwis provides two options for sorting the output (chronologically or by frequency). The default is sorting the hits by frequency of the translation variants. Figure 1 shows the top results when querying the alpine texts for the German word *Hütte* (engl. cabin). The corpus contains 744 sentences where *Hütte* was translated with the French word *cabane* and 217 sentences where it was translated with *refuge*. Pie charts indicate the percentage of each translation variant with respect to all translation hits for a given query.

All sentences that match the query are displayed, together with their source information (year, article number, source language, title and author) and their aligned sentences (translations). The search word and its aligned word or phrase are highlighted.

Our German corpora provide for a special search feature. When searching lemmas, one wants to find all forms of the lemma. This is particularly tricky when word forms are split as is the case with German verbs that can have separated prefixes. For example, the verb *anfangen* (to begin) will be split into *fängt* + *an* in a German matrix clause. In between the verb stem and the separated prefix there are typically objects and adjuncts, which means that the two parts of the verbs can be many tokens apart.

When we prepared our German corpora for Bilingwis we assigned PoS tags to all tokens. The PoS tagger also assigns lemmas to each token that it knows from its training

[6] see https://www.tausdata.org

[7] see http://www.parasol.unibe.ch

[8] see http://corpus.leeds.ac.uk/paraquery.html

[9] see http://opus.lingfil.uu.se/lex.php

[10] see http://kitt.cl.uzh.ch/kitt/bilingwis/

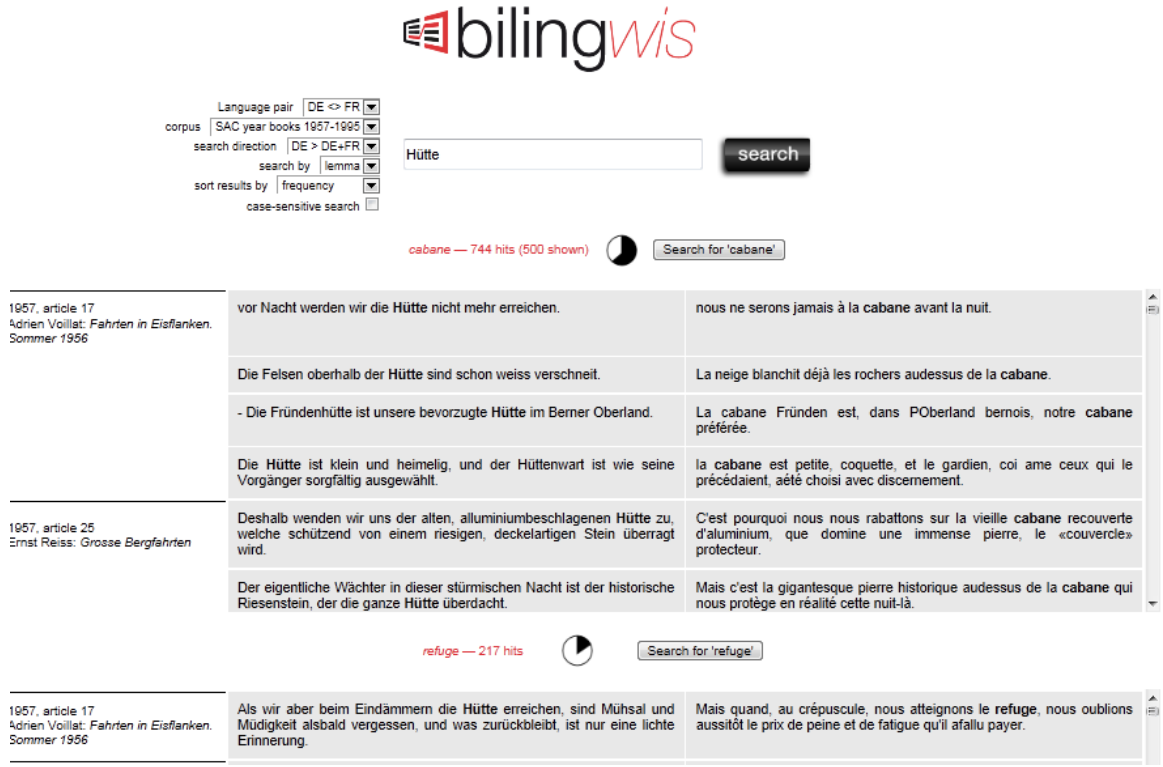

Figure 1: Bilingwis query for German *Hütte* with translation examples for *cabane* and *refuge*

corpus. Unfortunately the PoS tagger is not able to recombine the separated verb prefix with the verb stem. Therefore we employ a special-purpose re-attachment tool after PoS tagging. If the tool finds a token that is tagged as finite verb (e.g. *fängt* with lemma *fangen*) and later in the same sentence finds a token tagged as separated prefix (e.g. *an*), it preposes the prefix to the lemma of the verb stem. This operation is subject to a sanity check whether the prefixed verb actually is a valid German verb. We only approve the prefixed verb if it occurs as such in the corpus.
Preparing the corpus in this way allows Bilingwis to find both separated and attached forms of the same verb with a lemma query. For example, a query for *anfangen* will find the following examples with their French counterparts.

(1) DE: Auf ca. 3300 m **fängt** die Zone der hochgelegenen Weiden und diejenige der überaus langen Moränen **an**.
FR: A 3300 m environ **commence** la zone des hauts pâturages et celle des très longues moraines.

(2) DE: **Angefangen** hat der Spitzenkletterer seine Laufbahn vor ca. neun Jahren, als er mit seinen sportbegeisterten Eltern klettern ging.
FR: Son parcours a **commencé** il y a neuf ans environ lorsqu'il a découvert la grimpe avec ses parents.

Eight out of 21 occurrences of the verb *anfangen* in our alpine corpus are in the separated form including the above example sentence with 11 tokens between the verb and the prefix. To our knowledge Bilingwis is the only online search system over parallel corpora that finds verbs with separated prefixes.

While Bilingwis and all other online query tools for parallel corpora have only limited search options, we derive inspiration from full fledged query languages over parallel treebanks as in the TreeAligner. This is the kind of query language that linguists need for investigating the full potential of parallel corpora.

4.5. TreeAligner - Our Search Tool for Parallel Treebanks

The combined research on treebanks and parallel corpora has led to high-quality parallel treebanks (Volk et al., 2011b). For the purpose of aligning and searching such treebanks we developed the **TreeAligner** (Lundborg et al., 2007). This program comes with a graphical user interface to insert (or correct) word and phrase alignments between pairs of syntax trees. Figure 2 shows an example of a tree pair with word and phrase alignment lines. The lines denote translation equivalences. Both trees are constituent structure trees, but the edge labels contain function labels (like subject, object, attribute).
The TreeAligner is both an alignment tool and a powerful search tool over parallel treebanks (Volk et al., 2007, Marek et al., 2008). We based our design of the query language on the popular monolingual treebank tool TIGER-Search[11]. We re-implemented the TIGER-Search query language for the TreeAligner. This language allows the user to query not only for lexical features (word form, lemma, PoS tag, morpho-syntactic features like case and gender) but also for dominance relations within the trees. In the TreeAligner this functionality is available for formulating conditions

[11] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch

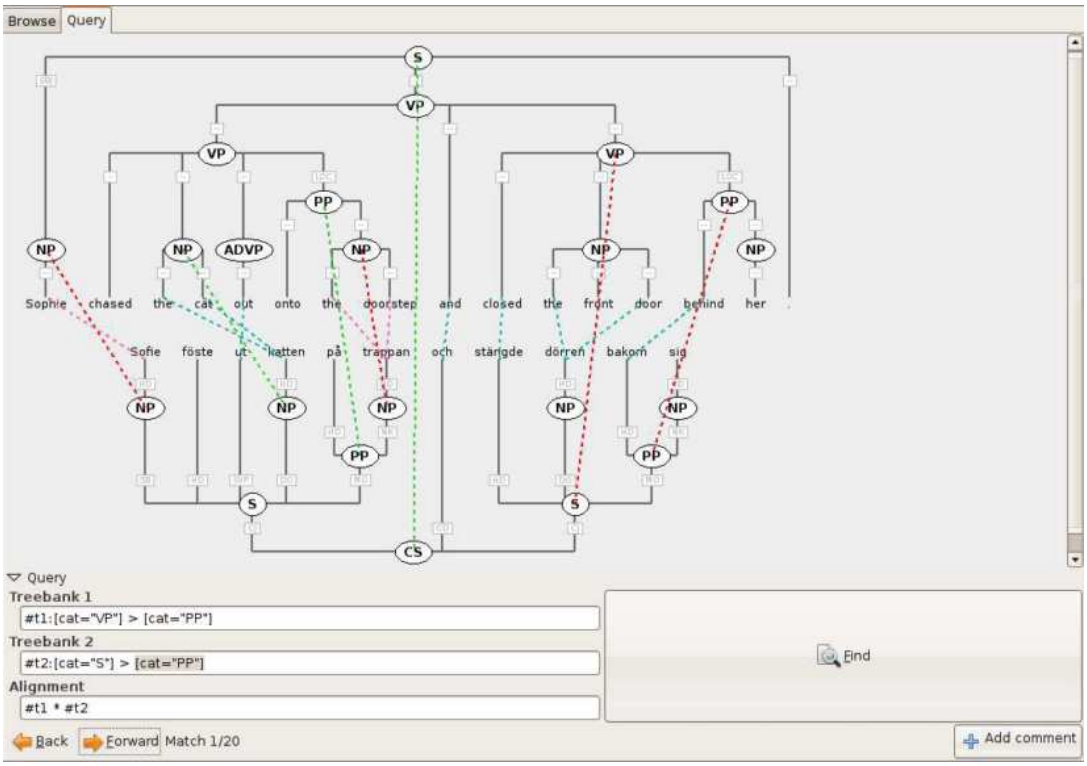

Figure 2: Example query over an English-Swedish Parallel Treebank in the TreeAligner

over both trees. In addition, the user can make bilingual queries by adding alignment conditions of the type "Find a phrase X in German WHEN it is aligned with a phrase Y in English" (cf. (Volk et al., 2007)). The query in figure 2 states: "Find a sentence with a verb phrase VP in Treebank 1 (i.e. the English treebank) which dominates a PP, AND find a sentence S in Treebank 2 (i.e. the Swedish treebank) which dominates a PP, WHEN the VP node and the S node are aligned across the languages.

Note that the TreeAligner query language comes with universal quantification which allows an intuitive interpretation of negated searches "Find a phrase X that does not contain word Y". This is very useful for linguistic investigations. Currently none of the above search systems over parallel corpora allows bilateral constraints nor negated queries.

We are working on filling this void within the recently started SPARCLING project ("Searching large PARallel Corpora for LINGuistics"). As a first step we have designed and implemented a database to store large corpora. We have cleaned Europarl for better recall and aligned it on the paragraph and word level. We are now in the process of porting the TreeAligner query language to our new search system.

5. Conclusions

We have shown that online parallel corpus query systems have become popular recently. We have argued that they fall into three distinct classes. First, there are systems for language learners and the interested public, second there are systems for translators that allow searches in domain-specific corpora, and third there are systems for researchers working in contrastive linguistics and translation science. Interestingly they show diagonal developments. The general public systems (like Glosbe, Linguee, Tradooit) employ automatic word alignment while the systems for the researchers (like ParaSol and ParaQuery) have not tapped the potential of this new technology.

We believe that research-oriented query systems over large parallel corpora need a powerful query language like the one in the TreeAligner, a tool for searching small to medium parallel treebanks. This language must include queries with conditions over both sides of a parallel text plus conditions on the word and phrase alignment. The language will be even more useful if it includes full negation constraints.

Parallel corpora offer a wealth of linguistic information. Some companies have started to exploit parallel corpora for enriching online dictionaries and providing popular services. Now researchers in linguistics and translation science begin to realize the added value of large collections of aligned translated texts. After all translations provide a sort of annotation and thus enable search precision that is not possible on monolingual corpora.

Acknowledgments

We would like to thank Sciex for grant 11.002, which supported our Dagstuhl workshop on "Annotation and Alignment of Parallel Corpora for Linguistic Research". We also gratefully acknowledge support by the Swiss National Science Foundation for project 105215_146781/1 on "Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation".

6. References

Karin Aijmer and Bengt Altenberg, editors. 2013. *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson*, volume 54 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam.

Michael Barlow. 2002. ParaConc: Concordance software for multilingual parallel corpora. In *Proceedings of LREC. Third International Conference on Language Resources and Evaluation*, Las Palmas.

Lars Borin, editor. 2002. *Parallel Corpora, Parallel Worlds. Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999*, volume 43 of *Language and Computers*. Rodopi, Amsterdam.

Julien Bourdaillet, Stéphane Huet, Philippe Langlais, and Guy Lapalme. 2010. Transsearch: from a bilingual concordancer to a translation finder. *Machine Translation*, 24:241–271.

Lindsay Bywood, Martin Volk, Mark Fishel, and Yota Georgakopoulou. 2013. Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives: Studies in Translatology*.

Bruno Cartoni and Marie-Aude Lefer. 2011. Negation and lexical morphology across languages: insights from a trilingual translation corpus. In J. Fernandez-Dominguez, M.-A. Lefer, and V. Renner, editors, *Poznan Studies in Contemporary Linguistics, special issue on English Word-Formation in Contrast*.

Jarle Ebeling. 1998. The Translation Corpus Explorer: A browser for parallel texts. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*, volume 24, pages 101–112. Rodopi.

Knut Hofland and Stig Johansson. 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, pages 87–100. Rodopi, Amsterdam.

Stig Johansson. 2007. *Seeing Through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*, volume 26 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam.

Ann Lawson. 2001. Collecting, aligning and analysing parallel corpora. In Mohsen Ghadessy, Alexander Henry, and Robert L. Roseberry, editors, *Small Corpus Studies and ELT. Theory and Practice*, volume 5 of *Studies in Corpus Linguistics*, pages 279–309. John Benjamins, Amsterdam.

Joakim Lundborg, Torsten Marek, Maël Mettler, and Martin Volk. 2007. Using the Stockholm TreeAligner. In *Proc. of The 6th Workshop on Treebanks and Linguistic Theories*, Bergen, December.

Torsten Marek, Joakim Lundborg, and Martin Volk. 2008. Extending the TIGER query language with universal quantification. In *Proceeding of KONVENS*, pages 3–14, Berlin.

Michael P. Oakes and Meng Ji, editors. 2012. *Quantitative Methods in Corpus-Based Translation Studies. A practical Guide to Descriptive Translation Research*. Number 51 in Studies in Corpus Linguistics. John Benjamins.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Wolfgang Teubert. 1996. Comparable or parallel corpora. *International Journal of Lexicography (Special Issue)*, 9(3):238–264.

Jörg Tiedemann. 2011. *Bitext Alignment*. Morgan & Claypool Publishers.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, Istanbul.

Martin Volk, Joakim Lundborg, and Maël Mettler. 2007. A search tool for parallel treebanks. In *Proc. of Workshop on Linguistic Annotation at ACL*, pages 85–92, Prague.

Martin Volk, Anne Göhring, Stéphanie Lehner, Annette Rios, Rico Sennrich, and Heli Uibo. 2011a. Word-aligned parallel text. A new resource for contrastive language studies. In *Proceedings of SDH 2011 - Supporting Digital Humanities: Answering the Unaskable*, Copenhagen.

Martin Volk, Torsten Marek, and Yvonne Samuelsson. 2011b. Building and querying parallel treebanks. *Translation: Computation, Corpora, Cognition (Special Issue on Parallel Corpora: Annotation, Exploitation and Evaluation)*, 1(1):7–28.

Ruprecht von Waldenfels. 2011. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In M. Daniela and R. Garabik, editors, *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011*, pages 156–162, Bratislava.

Manuela Weibel. 2014. Aufbau paralleler Korpora und Implementierung eines wortalignierten Suchsystems für Deutsch - Rumantsch Grischun. Master thesis, University of Zurich, Institute of Computational Linguistics.