Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian languages

Hercules Dalianis Martin Rimka Viggo Kann *
Dept of Computer and System Sciences, School of Computer Science and Communication*
KTH and Stockholm University
Forum 100, 164 40 Kista, Sweden
Email: hercules@dsv.su.se, rimka@dsv.su.se, viggo@csc.kth.se

Abstract

This paper presents how we adapted a website search engine for cross language information retrieval, using the Uplug word alignment tool for parallel corpora. We first studied the monolingual search queries posed by the visitors of the website of the Nordic council containing six different languages. In order to compare how well different types of bilingual dictionaries covered the most common queries and terms on the website we tried a collection of ordinary bilingual dictionaries, a small manually constructed trilingual dictionary and an automatically constructed trilingual dictionary, constructed from the news corpus in the website using Uplug. The precision and recall of the automatically constructed Swedish-English dictionary using Uplug were 71 and 93 percent, respectively. We found that precision and recall increase significantly in samples with high word frequency, but we could not confirm that POS-tags improve precision. The collection of ordinary dictionaries, consisting of about 200 000 words, only cover half of the top 100 search queries at the website. The automatically built trilingual dictionary combined with the small manually built trilingual dictionary consists of about 2000 words and covers 27 of the top 100 search queries.

Key words: Cross language information retrieval, parallel corpora, word alignment, Swedish, Danish, Norwegian.

1 Introduction

Scandinavian languages as Swedish, Norwegian, and Danish are comprehensible for Scandinavians. A typical Swede will for example understand written and to a certain degree spoken Danish, but is not able to speak Danish, that is he has a passive understanding of Danish (and vice versa for the other speakers).

The development of Internet has caused a new problem: the Scandinavians have difficulty finding information in the other neighboring languages since they do not have active knowledge in the other languages and therefore cannot write correct search queries.

The Nordic council experiences exactly such a problem on its website http://www.norden.org, since it has information in the main Nordic languages: Swedish, Danish, Norwegian, Icelandic, Finnish as well as English. The three languages Swedish, Danish and Norwegian are by the Nordic council considered to be one language – Scandinavian – and intercomprehensible, and are therefore not translated into their counterparts. Both employed and visitors at the website have difficulty finding information since the information in the Scandinavian languages are not overlapping and the users are not active users of two or more of the Scandinavian languages. The Nordic council therefore sponsored a research project to construct a Nordic on-line dictionary (Kann & Hollman 2007) and a cross language search engine to make it possible to search in for example Swedish and also find information in Danish and Norwegian. The research presented in this paper was done in this project.

2 Previous research

Most approaches to cross language information retrieval use general bilingual dictionaries, for

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 26–33

example Indonesian-English, the MUST system, (Lin 1999) Amharic-English, CLEF, (Argaw et al 2004), Chinese-Japanese-English-Spanish-German, web search engine, (Zhou et al 2005), French-English, Questioning answer system (Plamondon & Foster 2003). One interesting approach in cross language information retrieval is the approach in Järvelin et al (2006) using fuzzy matching as the only translation technique for the two closely related languages Swedish and Norwegian.

There is a lack of bilingual dictionaries between small languages. A solution would be to use existing bilingual dictionaries between a small and a large language to create a bilingual dictionary for two small languages. This method is called pivot alignment and is argued for in Borin (2000). Borin writes that "Pivot alignment in-creases word alignment recall, without sacrificing precision", but in Zhou et al (2004) pivot language translation is said to make a 52% drop in performance compared to direct translation.

Charitakis (2007) used Uplug for aligning words in a Greek-English parallel corpus. The corpus was comparably sparse and unannotated, containing 200 000 words from each language downloaded from two different real bilingual websites. A sample of 498 word-pairs from Uplug were evaluated by expert evaluators and the result was 51 percent correct translated terms (frequency >3). When studying high frequent word pairs (>11), there were 67 percent correct translated terms. Velupillai & Dalianis (2008) showed 94 percent correct translation (in average) on the closely related languages Swedish, Danish and Norwegian using Uplug.

The ITools suite for word alignment was used in Nyström et al (2006) on a medical parallel corpus containing 174 000 Swedish words and 153 000 English words, thereby creating 31 000 terms with 76 percent precision and 77 percent recall.

It is well known that stemming in information retrieval increases precision and recall (e.g. Carlberger et al 2001), therefore one could assume that stemming eventually would improve word alignment. However, Strömbäck (2005) has experimented to use lemmatization before executing Uplug on an English-Swedish corpus, and his results do not give any clear indication whether stemming is useful in word alignment.

Schrader (2004) shows that lemmatization and tagging of English and German parallel text decrease precision but improve recall in word alignment.

Toutanova et al (2002) showed up to 16 percent error reduction in word alignment for English and French (Hansard parallel corpora) using POS tagging.

Compound splitting, which can be done automatically with high accuracy (Sjöbergh and Kann 2006), is another approach that could give good results before performing word alignment, see Popoviç et al (2006), though they do not write how large the improvement is.

Thus, the previous research raised a number of important research questions and problems: Does POS-tagging improve word alignment quality? What is the optimal size of the parallel corpus to obtain good quality bilingual dictionaries? Is lemmatization or stemming before word alignment a good approach to increase precision/recall? How useful is a pivot language in the process of creating bilingual dictionaries, and what is the best pivot language to use in this project? What is the lowest word frequency for a good quality word alignment?

3 Content of website and search behavior

The website experimented on was the website of the Nordic council containing around 40 000 web pages written in six different languages. To find out the search behavior of the users and also find out what type of information (and in which languages) is available at the website of the Nordic council, we connected the commercial search engine SiteSeeker and its search box to the Nordic council's web site and let the search engine run for six months. By this experiment we found the most common search queries, the search queries with no answers, in which languages the queries were written, etc.

Around 10 000 search queries are made per month on the website. The queries are in many different languages, most often in Swedish, English and Finnish.

Very early we took the 100 most common search queries posed to the website of the Nordic council and translated them manually to the other Scandinavian languages, i.e. manually cre-

ated and customized a Scandinavian dictionary. When we later got better statistics of the search queries we found that this trilingual dictionary in fact only covers 24 of the 100 most common search queries.

From the website we also extracted from each of four languages the 200 words with the highest tf-idf, that is the most significant words in each language on the website. These 800 words hence gave us a picture of the website.

We compared these words with a collection of bi- and trilingual dictionaries that we had access to, to find the coverage of the dictionaries. The dictionaries were the Lexin dictionaries Swedish-English, English-Swedish, Danish-Swedish, and Norwegian-Swedish-English, and the Nordic council Skandinavisk ordbok which is Swedish-Danish-Norwegian. The dictionaries contain altogether about 200 000 unique words. We found that of the 200 most common terms in each language on the website, on average 73 percent were covered by these dictionaries. The manual dictionary of 231 words covered 9 percent of the 800 most common search words on the website and 24 percent of the 100 most common search queries.

The collection of dictionaries covered only half (54) of the 100 most common search queries. It was reassuring to see that the entire website covered 98 of the 100 most common search queries (in practice 100 percent, since the only uncovered search queries "indtaste søgeord" and "skrifið leitarorð", meaning "Enter search words", were predefined queries at the website).

In order to be really useful for cross language searching the bi- and trilingual dictionaries have to be extended to all four languages (Danish, Norwegian, Swedish, and English). Even if this was done the amount of covered most common queries would probably still be about half.

Dalianis (2002) showed that one cannot use ordinary dictionaries for good quality automatic spell checking of queries to search engines. Ordinary dictionaries do not really match the very domain specific content on a website. Our covering results confirm this.

4 Corpora

The covering analysis motivated us to automatically build a trilingual dictionary using parallel news texts from the Nordic council website.

The news texts are mostly written in one language and then translated to three other languages, so that each article will exist in English, Finnish, Icelandic, and Scandinavian. Swedish, Danish, and Norwegian are thus considered to be one language, and therefore news written in one of these languages is not translated to the other Scandinavian languages. For example, a news text written in Swedish is translated into English, Finnish, and Icelandic, but not to Danish or Norwegian.

The consequence of this is that English, Icelandic, and Finnish can be considered to be pivot languages for Swedish, Danish, and Norwegian.

We extracted 4 873 news articles in RSS format, written in Swedish, Danish, Norwegian, and English. These articles were comparably short, in average containing 160 words per article, in total 260 000 words per language, except for English where there were 865 000 words, see table 1. Each English version of a news article had always a parallel version written in either Swedish, Danish, or Norwegian.

Parallel texts	No of news texts	English words	Swe/Dan/Nor words
Eng-Swe	1 569	259 364	229 215
Eng-Dan	1 638	299 992	272 516
Eng-Nor	1 666	305 866	278 626
Total	4 873	865 222	780 357

Table 1. Number of news texts and words in different corpora

Apart from the news texts, the Nordic Council website contains other parallel or semi-parallel texts, for example organization, regulations, procedures, fact sheets etc. However, these documents are very few compared to the news texts.

5 Word alignment

As a word alignment tool we decided to use Uplug, since many researchers recommended it and Uplug has been used with successful results for other languages, e.g. Swedish and Turkish (Megyesi & Dahlqvist 2007).

Uplug is a word alignment tool for parallel corpora and was developed at Uppsala University by Jörg Tiedemann (Tiedemann 2003, Uplug 2008). Uplug works excellent (we have used version 0.1.9d) even though it can be memory consuming, mostly when doing sentence alignment in large corpora. The memory problem, however, can be easily solved with 'hard delimiter' tags (Gale and Church 1991).

We executed Uplug on the parallel texts written in English and Swedish, English and Danish, and English and Norwegian.
The news articles were extracted from the RSS file, language classified with LingPipe (2006), and merged into one corpus file per language. To allow sentence alignment only within article boundaries, we added hard delimiters.

The corpus files were tokenized with built-in Uplug scripts and aligned with a sentence aligner based on the statistical model of sentence length (Gale and Church 1991). The output was then word aligned with Uplug, which uses a combination of statistical and linguistic information to align single and multi-word units (Tiedemann 2003). The Uplug output was presented both in XML format (with word link certainty and other clues) and in text format, as a frequency table with word frequency, source and target terms (table 2).

40	sustainable	hållbar
40	responsibility	ansvar
40	proposal	förslag
40	increase	öka

Table 2. English-Swedish frequency table

According to rough manual estimation, word links with frequency 3 and higher had much better precision than links with low frequency (1-2).

We also executed Uplug on corpora that were lemmatized with CST Lemmatiser (Jongejan and Haltrup 2005); however, we could not see any significant improvement in the Uplug output. We attributed this fact to insufficient accuracy in the lemmatization rules, and thus continued to use corpora with inflected forms remaining. The English-Swedish, English-Danish, and English-Norwegian frequency tables were used to create a Swedish-Danish-Norwegian dictionary using English as pivot language (Borin 2000, Sjöbergh 2005). The Swedish, Danish, and Norwegian tokens which were linked to the identical English tokens were considered to be equivalents. For example, Swedish *hållbar*, Danish *bæredygtig*, and Norwegian *bærekraftig* were linked in the Uplug output to the English word *sustainable* (table 3); therefore the three Scandinavian words could be aligned to each other.

This method is rather approximate and may align words which do not have the same meaning. Nevertheless, we found it useful in creating multi-lingual dictionaries for expanding search queries. To achieve better precision, we extracted only links with frequency 3 or above.

Frequency table	Word link	
Eng-Swe	*sustainable*	*hållbar*
Eng-Dan	*sustainable*	*bæredygtig*
Eng-Nor	*sustainable*	*bærekraftig*

Table 3. Example with Swedish, Danish, and Norwegian tokens aligned to an English token

One spin-off effect of such pivot alignment method was that we obtained synonym lists in each of the aligned languages. For example, if English *production* was linked to Swedish *produktion* and *tillverkning*, then both Swedish words could be considered synonyms and obtained using the same software as for extracting Scandinavian triplets. The same method was used by Kann and Rosell (2005) constructing possible synonym pairs that were later evaluated by Internet users.

Coverage	200 000 words in dictionaries	231 words in manual dictionary	1984 words in half-automatic dictionary	Complete website
800 most common words on website	76 %	9%	24%	100%
100 most common search queries	54 %	24%	27%	98%
250 most common search queries	36 %	14%	17%	98%

Table 4. Coverage of the website and queries by dictionaries

For production purposes, we obtained 805 triplets in Swedish-Danish-Norwegian (1834 unique words), from Uplug results and after pivot alignment that later were manually corrected (half-automatic dictionary) and merged with the manually constructed trilingual dictionary. This merged dictionary containing 1984 unique words was integrated in the SiteSeeker search engine to support the cross-lingual information retrieval on the Nordic council website. We investigated how this half-automatic dictionary covers the common words and queries of the website of the Nordic council. The coverage is about half of that for the 100 times larger collection of dictionaries, and it is more useful for cross-language searching, since it is not just bilingual. Table 4 summarizes the coverage results for evaluation purpose. We aligned the Swedish and English corpus with and without part-of-speech (POS) tags. The corpus was tagged using the TNT tagger (Brants 2000). The English model was trained on the Penn Treebank corpus. The Swedish model was trained on the Stockholm-Umeå Corpus (SUC) annotated with the Parole tagset (Megyesi 2001).

6 Evaluation

To evaluate the Uplug output, we used a prior evaluation method with gold standards (Ahrenberg et al 2000). This evaluation requires additional tailor-made software. However, one can re-use the gold standards for different types of parallel corpora (e.g. with and without POS-tags). In addition, prior evaluation allows for more accurate measurement of the system output because it is based on the corpora used by the system.

The gold standards were built by manually annotating links in the sentence-aligned Swedish-English parallel corpora, in accordance to the manual annotation guidelines (Merkel 1999). We omitted, however, the definite articles in the gold standards in order to make them more consistent with the bilingual lexicons required for the query expansion. The articles and other stop words are not included in such lexicons because these words have low significance in normal search.

To build the gold standard, we used a sample of the 5 000 most frequent search queries from the Nordic council website. We chose this type of sample in order to examine how the extracted bilingual lexicon can support the query expansion in parallel corpora.

We established that 647 terms (13% of the sample) could be found in the Swedish corpus used by Uplug in word alignment. These terms were divided into three frequency categories (table 5). The terms from each frequency category were then used to build a separate gold standard. The fourth gold standard was built by merging the first three gold standards, i.e. it contained terms from all frequency categories (337 terms).

We intended to make the gold standards as extensive as possible, but we also applied certain limitations on the sample to make it more close to the bilingual dictionary needed to support query expansion. Thus, the gold standards included only Swedish nouns and adjectives with different spelling than their English equivalents. The words with identical spelling as their translations (most of the proper names and abbreviations) were omitted because they did not require query expansion, and hence, were not important for evaluation. The sample terms with missing or indirect translations were also left out, i.e. only 'regular' links were allowed in the gold standards.

Frequency category	Sample terms found in Swedish corpus	Sample terms included in gold standards
1-2	229	91
3-10	206	111
>10	212	138

Table 5. Distribution of sample terms across frequency categories

The evaluation was done with the built-in Uplug script *evalalign.pl* which uses the MWU measures (Tiedemann 2003). These measures are tailored to produce more reliable values for precision and recall in the system links which contain multi-word units (MWU).

Table 6 presents precision values for the Swedish-English corpora measured against the four gold standards. We evaluated word alignment in the two types of Swedish-English corpora – without linguistic information (default pre-processing) and with it (POS-tags).

The main purpose of this evaluation was to measure the quality of Uplug used on the Nordic council corpus. We also wanted to examine whether POS-tags can improve word alignment.

Frequency category	Corpora with default pre-processing	Corpora with POS-tags
1-2	54%	54%
3-10	70%	67%
>10	83%	76%
all freq	**71%**	67%

Table 6. Precision in the Swedish-English corpora

Several conclusions can be made from this table. First, not surprisingly, words with higher frequency are aligned with better precision. For example, rare words which occur only once or twice in Swedish corpus show 54% precision, whereas words with frequency above 10 have 83% precision. These results are also very close to the results of Strömbäck (2005).

Next, the gold standard based on the middle frequency category (3-10) returns similar precision value as the gold standard consisting of terms in all frequency categories. In other words, the middle category is representative of all frequency categories together.

These two observations are consistent across both the default and POS-tagged corpora.

Finally, precision of the POS-tagged corpora in all frequencies (67%) is lower than precision of the corpora without POS-tags (71%). We can also observe that the difference between the default and POS-tagged corpus increases in middle and high frequency categories. Thus, the lowest frequency category shows almost identical precision for both types of corpora, whereas the difference between the precision values in the highest frequency category reaches 7%.

Frequency category	Corpora with default pre-processing	Corpora with POS-tags
1-2	82%	83%
3-10	95%	92%
>10	98%	96%
all freq	**93%**	91%

Table 7. Recall in the Swedish-English corpora

Table 7 presents recall values for the Swedish-English corpora. In this table, we can observe similar tendency across the recall values – the words with high frequency produce better recall values compared to the words with low frequency. Furthermore, the corpus with POS-tags has lower recall value than the corpus without POS-tags, except for the lowest frequency category.

On the other hand, the difference among the recall values in the default and POS-tagged corpus is not as distinct as among the precision values.

7 SiteSeeker uses bilingual dictionaries

The cross language dictionary with the 805 triplets in Swedish, Danish and Norwegian was connected to the SiteSeeker search engine. The search works as a query expansion expanding the original term to terms in the others languages provided the original term has a translation to another term. The interface can filter the hit lists based on language, see figure 1. 30 percent of the top 100 queries used cross-lingual information retrieval. The top 100 queries compose 8 percent of the total queries, and the top 5 000 queries compose 50 percent of the total queries. Of the top 100 queries 24 percent were proper nouns that of course were not translated.

Figure 1 shows an example of the cross language search on the Nordic council website. The Swedish word *arbetsmarknad* in the original search query *nordisk arbetsmarknad* is expanded to the Danish word *arbejdsmarked* which allows retrieving the relevant documents in Danish.

During 2006, the search statistics of SiteSeeker showed 36 percent queries with no hits. During 2008, with the cross language dictionary connected to SiteSeeker, we obtained only 19 percent queries with no hits, about half of the 2006 value, even though the site had about the same amount of indexed pages as in 2006.

8 Conclusions

Our conclusions from the experiments with the website of the Nordic council are that it is very difficult to obtain a large enough parallel corpus to automatically create a large enough bilingual or trilingual dictionary covering all types of queries from the users.

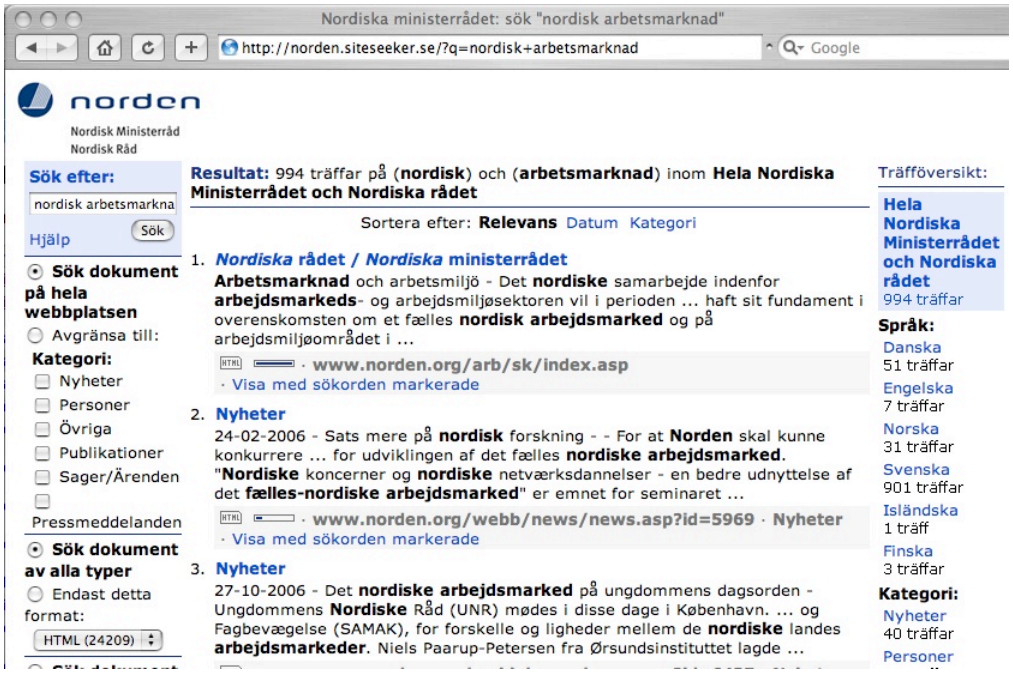

Figure 1. Cross language search on the Nordic council website

In order to improve the coverage a supplementary trilingual dictionary could be manually built using statistics of the top queries.

Word alignment quality using Uplug was high considering the small corpus. Also, we discovered that POS-tagging did not improve word alignment.

Pivot alignment is a useful trick that made our work possible. The similarity between the Scandinavian languages made the drop in performance due to the pivot alignment too small to be visible.

We post-processed the dictionary removing duplicate translations and translations that contained words that were shorter than four characters. This increased the quality and usefulness of the trilingual dictionary considerably.

The extracted words of the 4 873 news texts did not really cover the words in the 40 000 web pages, but when combined with a small hand-made trilingual dictionary they covered the most common search queries reasonably well.

Future work will encompass the impact of lemmatization in word alignment and as well as the use of other word alignment tools.

References

Ahrenberg, L., M. Merkel, A. Sågvall Hein and J. Tiedemann. 2000. Evaluation of word alignment systems. In Proc. LREC 2000, Athens.

Argaw, A., L. Asker, R. Cöster and J. Karlgren 2004. Dictionary-based Amharic - English Information Retrieval. In Proc. Cross Language Evaluation Forum (CLEF 2004), Bath, UK.

Borin L. 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In Proc. 18th International Conference on Computational Linguistics. COLING 2000, Vol. 1. Saarbrücken: Universität des Saarlandes. 97-103.

Brants, T. 2000. TnT - A Statistical Part-of-Speech Tagger. In Proc. Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.

Carlberger, J., H. Dalianis, M. Hassel, and O. Knutsson 2001. Improving Precision in Information Retrieval for Swedish using Stemming. In Proc. Nodalida 2001, Uppsala, Sweden

Charitakis K. 2007. Using parallel corpora to create a Greek-English dictionary with Uplug. In Proc. Nodalida 2007, Tartu, Estonia.

Dalianis, H. 2002. Evaluating a Spelling Support in a Search Engine. In Proc. Natural Language Processing and Information Systems, 6th International Confer-

ence on Applications of Natural Language to Information Systems, NLDB 2002 (Eds.) B. Andersson, M. Bergholtz, P. Johannesson, Stockholm, Sweden, June 27-28, 2002. Lecture Notes in Computer Science. Vol. 2553. pp. 183-190. Springer Verlag.

Gale, W. A. and K. W. Church 1991. A program for aligning sentences in bilingual corpora. In Proc. 29th annual meeting on Association for Computational Linguistics, p.177-184, 1991, Berkeley, California.

Jongejan, B. and D. Haltrup 2005. The CST Lemmatiser. Center for Sprogteknologi, University of Copenhagen, version 2.9 (October 6, 2005), http://cst.dk/download/cstlemma/current/doc/

Järvelin, A., S. Kumpulainen, A. Pirkola and E. Sormunen 2006. Dictionary-independent translation in CLIR between closely related languages. In Proc. Dutch-Belgian Information Retrieval Workshop, TNO ICT, Delft, The Netherlands, 2006.

Kann, V. and J. Hollman 2007. Tvärslå – Defining an XML exchange format and then building an on-line Nordic dictionary. In Proc. Automatic Treatment of Multilinguality in Retrieval, Search and Lexicography, Workshop in Copenhagen 2007.

Kann, V. and M. Rosell 2005. Free construction of a Swedish dictionary of synonyms. In Proc. Nodalida 2005, Joensuu.

Lin, C-Y. 1999. Machine Translation for Information Access across Language Barrier: the MuST System. In Machine Translation Summit VII, Singapore.

LingPipe 2006. LingPipe is a suite of Java libraries for the linguistic analysis of human language, http://www.alias-i.com/lingpipe/

Megyesi, B. 2001. Data-Driven Methods for PoS tagging and Chunking of Swedish. In Proc. Nodalida 2001, Uppsala.

Megyesi, B. and B. Dahlqvist 2007. The Swedish-Turkish Parallel Corpus and Tools for its Creation. In Proc. Nodalida 2007, Tartu, Estonia.

Merkel, M. 1999. Annotation Style Guide for the PLUG Link Annotator. Technical report, Linköping University.

Nyström, M., M. Merkel, L. Ahrenberg, P. Zweigenbaum, H. Petersson and H. Åhlfeldt. 2006. Creating a medical English-Swedish dictionary using interactive word alignment in BMC medical informatics and decision making.

Plamondon, L. and G. Foster. 2003. Quantum, a French/English Cross-language Question Answering System. In Proc. Cross-Language Evaluation Forum (CLEF 2003), Trondheim.

Popović, M., D. Stein and H. Ney 2006. Statistical Machine Translation of German Compound Words. In Proc. FinTAL, 5th International Conference on Natural Language Processing, Springer Verlag, LNCS, pages 616-624, Turku.

Sarr, M. 2003. Improving precision and recall using a spell checker in a search engine. In Proc. Nodalida 2003, Reykjavik.

Schrader, B. 2004. Improving Word Alignment Quality Using Linguistic Knowledge. In Proc. International Conference on Language Re-sources and Evaluation, LREC 2004, Lissabon.

Sjöbergh, J. 2005. Creating a free digital Japanese-Swedish lexicon. In Proc. PACLING 2005, pages 296-300, Tokyo.

Sjöbergh, J. and V. Kann 2006. Vad kan statistik avslöja om svenska sammansättningar (What can statistics reveal about Swedish compounds), Språk och Stil 2006, vol. 16, pages 199-214.

Strömbäck, P. 2005. The Impact of Lemmatization in Word Alignment, Master thesis, Department of Linguistics and Philology, Uppsala University.

Tiedemann, J. 2003. Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing, Doctoral Thesis, Studia Linguistica Upsaliensia 1, ISSN 1652-1366, ISBN 91-554-5815-7.

Toutanova, K., H. T. Ilhan and C. D. Manning 2002. Extensions to HMM-based Statistical Word Alignment Models. In Proc. ACL Language Processing (EMNLP), Philadelphia, 2002, pp. 87-94.

Uplug 2008. Uplug is a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpus, http://uplug.sourceforge.net/

Velupillai, S. and Dalianis, H. (2008). Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. In Proc. 2nd MMIES Workshop: Multi-source, Multilingual Information Extraction and Summarization, held in conjunction with COLING-2008, Manchester.

Zhou, Y., J. Qin, H. Che and J. F. Nunamaker 2005. Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence. In Proc. 38th Hawaii International Conference on System Sciences 2005.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206