Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment

Yvonne Samuelsson and Martin Volk
Stockholm University
Department of Linguistics

Abstract

A parallel treebank consists of syntactically annotated sentences in two or more languages, taken from translated documents. These parallel sentences are linked through alignment. This paper explores the use of word n-gram alignment, computed for statistical machine translation, to create syntactic phrase alignment. We achieve a weighted $F_{0.5}$-score of over 65%.

1 Introduction

In recent years, the research areas of parallel corpora and treebanks have been combined into parallel treebanks. A parallel treebank is a collection of language data consisting of texts and their translations, which have been grammatically tagged and syntactically annotated. They also contain some kind of alignment information, links between the languages.

We are developing a German-English-Swedish parallel treebank, SMULTRON[1] (Stockholm MULtilingual TReebank), consisting of just over 1000 sentences in each language. The alignment has been drawn manually, on the sentence, phrase and word levels.

Previously, other researchers have done much work on sentence and word alignment and there are several methods for automatic alignment, but the alignment on phrase level has not been explored to the same extent. This paper looks at automating the alignment process, by using n-gram alignment computed for statistical machine translation (SMT) to create syntactic phrase alignment[2].

[1] See http://www.ling.su.se/dali/research/smultron/index.htm.

[2] Some of the experiments reported here where first conducted by the main author for the GSLT graduate course Statistical Methods, supervised by Joakim Nivre, in the spring of 2007.

De Smedt, K., Hajič, J. and Kübler, S. (Eds.)
Proceedings of the Sixth International Workshop
on Treebanks and Linguistic Theories (2007)

2 The SMULTRON parallel treebank

In creating the parallel treebank, we first annotated the monolingual treebanks with the ANNOTATE treebank editor[3]. We annotated the English treebank according to the Penn Treebank guidelines, (Bies et al., 1995), while the German follows the TIGER annotation schema, (Skut et al., 1997, Brants et al., 2002). For the Swedish treebank we used an adapted version of the German TIGER guidelines. These all give phrase structure (constituent) trees. Both the PoS tags and the syntactic structure were manually checked, and we automatically checked for completeness and consistency. An example tree can be seen in figure 1.

The TIGER annotation guidelines lead to flat trees. This means, for instance, no unary nodes, no "unnecessary" NPs (noun phrases) within PPs (prepositional phrases) and no finite VPs (verb phrases). This speeds up the annotation process, but we prefer to have "deep trees" to be able to draw the alignment on as many levels as possible. After completing the flat German and Swedish trees, we therefore automatically deepened the structures by inserting unambiguous nodes. This procedure has been described in (Samuelsson and Volk, 2004). The English guidelines lead to deeper trees, so they have not been automatically deepened.

The parallel treebanks consists of two parts, Jostein Gaarder's novel "Sophie's World", and economy texts. For the experiments with automatic alignment reported in this paper we only used the "Sophie" part of the parallel treebank. The English treebank contains 528 sentences (7,829 tokens and 7,020 non-terminal nodes) and the Swedish treebank 536 sentences (7,394 tokens and 5,351 nodes).

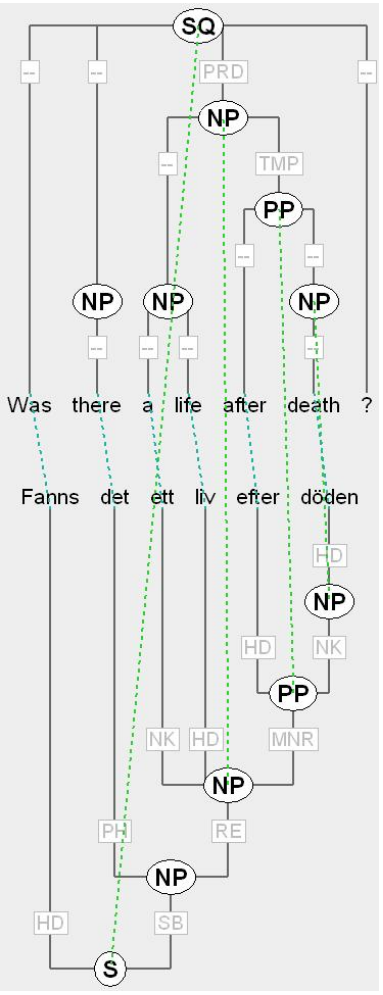

Figure 1: An example tree from our parallel treebank.

[3] www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

3 Alignment

Phrase alignment can be regarded as an additional layer of information on top of the syntax structure. It shows which part of a sentence in one language is equivalent to which part of a corresponding sentence in another language.

3.1 SMULTRON Alignment Guidelines

After creating the monolingual treebanks, we convert the trees into TIGER-XML, a powerful database-oriented representation for graph structures[4]. In a TIGER-XML graph each terminal node (token) and non-terminal node (linguistic constituent) has a unique identifier. We use these unique identifiers for the phrase and word alignment across trees in corresponding translation units. We also use an XML representation for storing this alignment.

We draw alignment lines manually between sentences, phrases and words over parallel trees. This is done with the help of our alignment tool, the Stockholm TreeAligner[5], a graphical user interface to insert (or correct) alignments between pairs of syntax trees. We want to align as many phrases as possible. The goal is to show translation equivalence. Phrases shall only be aligned if the tokens that they span represent the same meaning, if they could serve as translation units outside the current sentence context. The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit.

We differentiate between two types of alignment, displayed by different colours in our alignment tool, exact translation correspondence and fuzzy (approximate) translation correspondence. Our alignment guidelines allow phrase alignments within m:n sentence alignments. Even though m:n phrase alignments are technically possible, we have only used 1:n phrase alignments (not specifying the direction). The 1:n alignment option is not used if a node from one tree is realized twice in the corresponding tree.

Pronouns should not be aligned to full noun phrases. Nodes that contain extra information in one language should not be aligned. This means that e.g. a sentence (with the subject) cannot be aligned to a verb phrase (without the sentence subject).

Before we describe our experiments, let us take a quick look at some methods for automatic alignment of sentences, words and phrases.

[4] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html

[5] http://www.ling.su.se/dali/downloads/treealigner/index.htm. The latest version of the TreeAligner also has a query-function for parallel treebanks, see (Lundborg et al., 2007).

3.2 Automatic Sentence Alignment

There are well-established algorithms for aligning sentences (which represent translation correspondences) across parallel corpora. These algorithms are based on features such as sentence length (in terms of number of characters), word correspondences (as taken from bilingual dictionaries, or automatically found cognates), or distance factors. A well-known example is the algorithm presented by (Gale and Church, 1993) which uses length comparisons. It allows 1:n sentence alignments and optimizes the sentence mapping across paragraphs.

3.3 Automatic Word Alignment

Word and phrase alignment goes beyond sentence alignment in that it captures sub-sentential correspondences. Word alignment algorithms usually require sentence-aligned corpora. However, the features for computing word alignment cannot be the same as for sentence alignment. The word order is different across languages and length comparisons may help but are not indicative. Instead co-occurrence statistics can be used. If two words frequently co-occur in corresponding sentences, they are good candidates for translation correspondences. Because of e.g. different compounding dynamics in languages like English versus German, 1:n word correspondences must be applied.

Following (Tiedemann, 2003) we distinguish two types of word alignment approaches. He calls them association approaches and estimation approaches (they are also called heuristic models and statistical models by e.g. (Och and Ney, 2003)). Association approaches use string similarity measures, word order heuristics, or co-occurrence measures (e.g. mutual information scores). For the latter, the idea is to find out if a cross-language word pair co-occurs more often than could be expected from chance.

Estimation approaches, on the other hand, use probabilities estimated from parallel corpora, inspired from statistical machine translation, where the computation of word alignments is the basis of the computation of the translation model. The word correspondences computed by the freely available GIZA++ system (Och and Ney, 2000, Och and Ney, 2003) have constantly scored high in evaluations. All these methods include multi-word units as alignment targets, these units sometimes being precomputed and sometimes being determined during the alignment process.

3.4 Automatic Phrase Alignment

Sometimes it is not possible to establish correspondences on the word level; there are rather meaning equivalences on larger units. We capture this by using phrase alignments. For example, the co-ordinated phrase *die Papierindustrie und der Bausektor* certainly corresponds as a whole to *the pa-*

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

per and construction sector, but we would not want to align *Papierindustrie* to *paper* alone.

The term *phrase* is being used in different ways with regards to alignment, denoting both linguistic phrases and word n-grams of varying length. It is often claimed that using syntactic phrases does not improve SMT and small units, up to 3-grams, are sufficient for good accuracy (see e.g. (Koehn, Och, and Marcu, 2003)). In this paper we will use the word *phrase* in the linguistic sense, as the level between word and sentence (even though a phrase can consist of only one word or of a whole sentence), otherwise using the term n-gram, i.e. word n-grams.

There are two general approaches to phrase alignment (see e.g. (Schrader, 2007)), finding correspondences between phrases through parsing or chunking (based on e.g. co-occurrences), or deriving phrase alignment through previous word alignment. We explore the latter in this paper.

4 Tools for statistical machine translation

GIZA++ is an extension of the program GIZA (which is part of the SMT toolkit EGYPT). It is an implementation of the IBM Models (Brown et al., 1993), and was written by Franz Josef Och. The system computes word alignments between corresponding bilingual sentences according to statistical models. A detailed description of the software can be found in (Och and Ney, 2003).

Phillip Koehn's Pharaoh software (see (Koehn, 2004, Koehn, 2004)) is a decoder for SMT, but is available together with scripts for training the SMT system, including scripts for extracting n-grams from word-aligned sentences. The word alignments are taken from the intersection of bi-directional runs of GIZA++ plus some additional alignment points from the union of the two runs. From this a maximum likelihood lexical translation table is estimated. Pairs of n-grams are extracted that are consistent with the word alignment, meaning that an n-gram has to contain all alignment points for all its words. The end result is the translation model, the so called phrase-table (which we will continue to call phrase-table, even though it is an n-gram-table), which contains a set of multi-word unit correspondences in the form of a source n-gram, a target n-gram, and conditional probabilities. When referring to Pharaoh in this paper, we mean the scripts for extracting phrase-tables and not the decoder.

The Thot toolkit (Ortiz-Martínez, García-Varea, and Casacuberta, 2005) is for training phrase-based (n-gram) models for SMT. To calculate the n-gram alignments, we use the word alignment from GIZA++ as input. It is also possible to do operations between the word alignment matrices, to improve them, before extracting the phrase-based model. These are union (OR), intersection (AND), SUM and two different versions of symmetriza-

tion (symmetrization is described in (Och, 2002) and is a mixture of intersection and union).

5 Experiments

Our experiments basically have two phases. As can be seen in figure 2, first (after creating the treebanks and running GIZA++ on the texts to get the word alignment matrices) we used the SMT training modules Pharaoh and Thot to create the n-gram alignment. This was done for English-Swedish (EN-SV) and for Swedish-English (SV-EN). Thot was also used to produce combined word alignment matrices.

The output of each run of the SMT training modules is a phrase-table, which in the second phase is fed to the linguistic alignment filter program, together with the treebanks. This creates the alignment between the syntactic phrases of the parallel treebank, which is then evaluated against a manually created gold standard alignment-file. This gold standard also contains word alignment, which we ignored during evaluation, since the focus of the experiments is on phrase alignment.

5.1 The linguistic alignment filter

The program for creating the phrase alignment, the linguistic alignment filter, was written in Perl by the main author. It needs the TIGER-XML file for both language 1 (L1) and language 2 (L2), the phrase-table for L1-L2 and a file containing sentence alignment. The phrase-table contains the source language n-gram, the target language n-gram and the phrase translation probability for source|target.

The linguistic alignment filter program goes through the treebank file of one language and for each sentence extracts every phrase of the tree as a string containing the words that it spans. It then compares this string to the n-gram entries of the phrase-table. If there is a match between the string containing the phrase and the string containing the n-gram, the node number is stored together with the corresponding (aligned) n-gram (L2) and the probability. The number of possible links from the phrase-table is thus reduced to only the ones where the L1 n-gram equals a syntactic phrase. Then the second treebank is processed. Every node of every tree is again extracted as a string of words and this string is compared to the already stored L2 n-gram strings. If there is a match, the node identifier is stored. This reduces the number of alignment links further, only leaving links where both the L1 and the L2 strings equal syntactic phrases. In these experiments we do not distinguish between exact and approximate alignment.

Since the phrase-table does not contain any information about the context, if the n-gram *she* is aligned to the n-gram *hon* in the phrase-table,

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

every *she* in the English treebank would be aligned to every *hon* in the Swedish treebank. To avoid this, we used a separate XML-file with sentence alignment (automatically constructed but manually checked), to restrict the alignment links created. Only if there is a sentence link, there can be phrase links.

The phrase-table contains punctuation symbols, which are also part of the tree in the English annotation format. However, punctuation symbols are not included in the tree structure in the Swedish (or German) annotation format. Because of this, the symbols were removed. For concatenating the string of words of a syntactic phrase, a token not containing any alphanumerics was just not added to the string. Removing all punctuation symbols from the phrase-table string is not as simple, since we do not want to remove e.g. the apostrophe of *haven't*. The (rather simple) solution was to remove all punctuation symbols following or followed by a blank. Then, for tokenization, a blank is inserted before the remaining apostrophes, unless the apostrophe was surrounded by n and t, where the blank is inserted before the n. This might still induce some errors (e.g. the tokenization of *can't* should be *can - 't* and not *ca - n't*), but it will handle most cases.

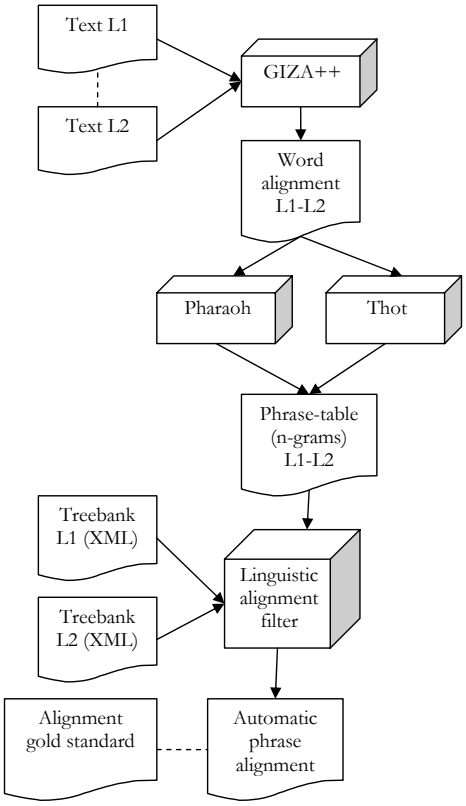

Figure 2: The basic experimental set-up.

5.2 Results

We used the whole Sophie treebank (528 English and 536 Swedish sentences) for the training run to extract the phrase-tables with both Pharaoh and Thot. The gold standard contains 3143 aligned node pairs (links). The calculated scores are precision, recall and a combined $F_{0.5}$-score, which weights precision twice as much as recall. We consider a high precision to be of more value, since we prefer to add links manually, with a minimal effort of manually correcting links.

In the following, when talking about alignment in a particular direction

Pharaoh		Phrase links	Precision	Recall	$F_{0.5}$-score
EN-SV	10-gram	1991	71.47%	45.28%	59.92%
	30-gram	2260	**73.10%**	52.56%	64.67%
	60-gram	2271	**73.10%**	52.82%	64.80%
SV-EN	10-gram	1964	72.40%	45.24%	60.33%
	30-gram	2230	**74.04%**	52.53%	65.15%
	60-gram	2239	74.01%	52.72%	**65.23%**
Thot		Phrase links	Precision	Recall	$F_{0.5}$-score
EN-SV	10-gram	2489	66.81%	52.91%	61.43%
	30-gram	2758	68.20%	59.85%	65.17%
	60-gram	2769	68.18%	**60.07%**	**65.25%**
SV-EN	10-gram	2532	62.60%	50.43%	57.94%
	30-gram	2803	64.50%	57.52%	62.00%
	60-gram	2813	64.56%	**57.78%**	62.13%

Table 1: Results for basic alignment after the linguistic alignment filtering.

(e.g. English-Swedish, EN-SV) we only refer to the direction of the phrase-table or the word alignment matrix, since it is constructed with one language as source and one as target. The phrase alignment of the parallel treebank, however, is not directional.

The results can be seen in table 1 and are shown for n-grams with a maximal length of 10, 30 and 60[6]. As expected, precision and recall increase with the length of the n-grams, even if the difference between 30 and 60 is negligible or even showing a drop in precision. Since a phrase can contain anything from one word to a whole sentence (the root node), longer n-grams match more phrases in the treebanks. However, longer n-grams also give larger phrase-tables. This is not a computational issue with only 500 sentences, but will be with a larger corpus[7]. The maximum number of tokens per sentence for our treebank (the sentence is the maximal length of a phrase) is 57 for English and 48 for Swedish. Pharaoh is slightly better with SV-EN than EN-SV, but they are basically the same. Thot, however, does much worse for SV-EN. This is the result of Pharaoh combining the uni-directional word alignment, which Thot does not. Comparing the two systems in general shows that Thot has higher recall (up to around 60%), while Pharaoh has higher precision (up to around 74%). The highest F-score

[6]When we talk about a phrase-table with an n-gram length of e.g. 10 this is the maximal length. Thus the phrase-table contains n-grams of lengths 1 to 10.

[7]As an example, the sizes of the Pharaoh phrase-tables are 942kB for 10-gram, 1864kB for 30-gram and 1947kB for 60-gram. To exemplify the reduction of alignment links through the filter program, the Pharaoh EN-SV 30-gram phrase-table contains 16,216 alignments, leaving 14% after filtering (see the number of phrase links in table 1), and the Thot EN-SV 30-gram phrase-table contains 34,267 alignments, leaving 8% after filtering.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

Pharaoh		Phrase links	Precision	Recall	$F_{0.5}$-score
merge	10-gram	1905	73.28%	44.42%	60.23%
intersect	30-gram	2167	**74.85%**	51.61%	65.08%
	60-gram	2176	74.82%	51.80%	**65.16%**
merge	10-gram	2050	70.68%	46.10%	60.02%
union	30-gram	2323	72.36%	53.48%	64.75%
	60-gram	2334	72.37%	**53.74%**	64.87%
Thot		Phrase links	Precision	Recall	$F_{0.5}$-score
merge	10-gram	1718	75.55%	41.30%	59.19%
intersect	30-gram	1963	77.08%	48.14%	64.21%
	60-gram	1971	77.07%	48.33%	64.32%
merge	10-gram	3303	59.04%	62.04%	60.01%
union	30-gram	3598	60.48%	69.23%	63.14%
	60-gram	3611	60.51%	69.52%	63.24%
AND	10-gram	3835	50.93%	62.14%	54.18%
	30-gram	4291	51.39%	70.16%	56.42%
	60-gram	4312	51.37%	**70.47%**	56.47%
OR	10-gram	1750	75.71%	42.16%	59.84%
	30-gram	1965	**77.15%**	48.23%	64.30%
	60-gram	1973	77.14%	48.43%	64.41%
SYM1	10-gram	2424	67.66%	52.18%	61.57%
	30-gram	2699	69.28%	59.50%	65.68%
	60-gram	2708	69.31%	59.72%	65.79%
SYM2	10-gram	2464	67.33%	52.78%	61.67%
	30-gram	2739	68.97%	60.10%	65.73%
	60-gram	2748	69.00%	60.32%	**65.84%**

Table 2: Results for bi-directional alignment after the linguistic alignment filtering.

is achieved by Thot for EN-SV with 60-grams.

Since there are apparent differences depending on which language is source and target, we merged the two phrase alignment files, once by taking the union of the links and once by taking the intersection. In addition to this, we combined the two uni-directional word alignment matrices with the help of Thot (through the AND, OR, SUM, SYM1 and SYM2 options) before creating additional phrase-tables. The results can be seen in table 2.

There are differences between Pharaoh and Thot with regards to the merged alignment. With Pharaoh, the F-scores are slightly better for the intersection than for the union. As expected, the union of the alignment links achieved a higher recall but a lower precision, while it is the other way around for the intersection of the links. Precision scores overall are better

than recall-scores. As expected, since Pharaoh already combines the uni-directional word alignment, the variations between intersection and union are only a few percentage points.

Thot also achieves a better F-score for intersection than for union, but only for larger n-grams, the F-score for 10-grams being higher for union. Again, precision is very high for intersection, while recall is very low. Recall is very high for union, while precision goes down. The differences between union and intersection are major.

The phrase alignment created through combining the word alignment matrices gave surprising results, since the merged intersection and the OR option gave similar results, as well as the merged union and the AND option. It is not clear why these combinations of the word alignment give results contradicting logic.

The AND option results in the highest recall of all the experiments, but it is still only slightly higher than for the merge union, while precision is much better for the merge union. The differences between merge intersection and the OR option are minor, even though the OR option gives slightly better both precision and recall. The precision for 30-grams with the OR option is the highest achieved in any of the experiments. The SUM option gave the same results as the OR option, since the differences only show in the probabilities of the phrase-tables, but not in the n-grams. The SYM options are very similar to each other and to the EN-SV uni-directional alignment, thus being better than the SV-EN alignment.

6 Conclusions

This paper is a report on experiments where freely available software for training statistical machine translation systems has been used to extract word n-gram alignments. These n-grams have been filtered to create syntactic phrase alignment.

The general trend is that phrase-tables containing longer n-grams give better results, but when including the full length of all sentences, precision tends to decrease slightly, with only minor improvements for recall. Pharaoh achieved precision of just under 75% and recall of over 50%, with Swedish-English being slightly better than English-Swedish. Thot achieved lower precision, around 65-68% but higher recall, around 60%, with English-Swedish being much better than Swedish-English.

In addition to the basic alignment, we experimented with combining the word alignment matrices (for Thot) and merging the final phrase alignment (for both Thot and Pharaoh). For Pharaoh, we would prefer using the intersection of the phrase alignment links. For Thot, it is not as clear, since the OR option gives the absolutely best precision, while the SYM options are more balanced between precision and recall and also achieve the highest

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

F-scores of all the experiments.

Since the Sophie-part of the treebank only contains around 500 sentences, the amount of training material for the automatic GIZA++-alignment is rather small. We conducted a first experiment, where the amount of training material was doubled by adding a file from Europarl (Koehn, 2002). Precision showed a major drop, while recall improved. Longer n-grams are generally too rare to be found several times. This means that the improvements from adding text would be in the probabilities for short n-grams. Further experiments with more text of the same type as the material to be aligned are needed.

There is another problem with frequent short n-grams. At the moment there is alignment between all instances of a word appearing multiple times in a sentence (which is often the case for e.g. pronouns). This is difficult to handle through statistical alignment.

Another problem with the filtering program is the fact that the (Swedish and German) treebanks allow for crossing edges. In the Swedish "Sophie" part of our treebank there are 85 crossing edges in 67 sentences. Since this gives us discontinuous constituents, these phrases cannot be matched to any n-grams, as long as n-grams are defined as adjacent words.

One possible improvement for the future would be to split unmatched phrases into pieces, and then to combine n-grams to find a match. One could start by splitting every phrase into two parts at every word boundary. If we for example have the phrase *her red shoe*, we could get `{<her><red shoe>}` and `{<her red><shoe>}`. We could then try to match both parts of L1 to the phrase-table and see if the n-grams of L2 can be combined into a matching phrase for L2. This is, however, not a trivial task.

In the future we would also like to do further experiments both with more languages (by adding German) and with more text types (by adding the Economy part of our parallel treebank).

References

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for treebank II style, Penn treebank project. Technical report, University of Pennsylvania.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol, Bulgaria.

Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Gale, William A. and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102.

Koehn, Philip. 2004. PHARAOH - a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models (User Manual and Description for Version 1.2). Technical report, USC Information Sciences Institute, August.

Koehn, Philipp. 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. December.

Koehn, Philipp. 2004. PHARAOH - Training Manual. Technical report, MIT CSAIL, July.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 127–133, Edmonton, Alberta, Canada, May 27 - June 1. ACL.

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

Och, Franz Josef and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proc. Of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Och, Franz Joseph. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, Computer Science Department, RWTH Aachen, Germany, October.

Ortiz-Martínez, Daniel, Ismael García-Varea, and Francisco Casacuberta. 2005. Thot: a Toolkit to Train Phrase-based Statistical Translation Models. In *Tenth Machine Translation Summit*, Phuket, Thailand, September. AAMT.

Samuelsson, Yvonne and Martin Volk. 2004. Automatic node insertion for treebank deepening. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany, December.

Schrader, Bettina. 2007. *Exploiting Linguistic and Statistical Knowledge in a Text Alignment System*. Ph.D. thesis, Universität Osnabrück.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 88–95, Washington, DC.

Tiedemann, Jörg. 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Uppsala university.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476