Approaching a New Language in Machine Translation

- Considerations in Choosing a Strategy

Anna Sågvall Hein, Per Weijnitz

Department of Linguistics and Philology, Uppsala University
Box 637, S-751 26 UPPSALA
anna@lingfil.uu.se, per.weijnitz@lingfil.uu.se

Abstract

As a contribution to the on-going discussions concerning what strategy to use when approaching a new language, we present our experience from working with Swedish in the rule-based and statistical paradigms. We outline the development of Convertus. a robust transfer-based system equipped with techniques for using partial analyses, external dictionaries, statistical models and fall-back strategies. We also present a number of experiments with statistical translation of Swedish involving several languages. We observe that the concrete language pair, translation direction and corpus characteristics have an impact on translation quality in terms of the BLEU score. In particular, we study the effects of the openness/closeness of the domain, and introduce the concept of corpus density to measure this aspect. Density is based on repetition and overlap of text segments, and it is demonstrated that density correlates with BLEU. We also compare a statistical versus a rule-based approach the translation of a Swedish corpus. The rule-based approach for which we use Convertus outperforms the statistical in a modest way. For both systems there is much room for improvement and it is likely that they both can be further developed to a BLEU score of 0.4 – 0.5 which seems good enough for post-editing to pay off. However, a major difference concerns the kinds of errors that are made and how they can be identified. The errors caused by Convertus can be easily traced and explained in linguistic terms and hence also avoided by extensions and modifications of the dictionaries and the grammars. The errors produced by the statistical system are, however, less predictable and difficult to pin-point and eliminate by further training. In particular, the many cases of omissions constitute a serious problem. Our conclusion will be that the investment made in developing a rule-based system, preferably backed up by a statistical system, will pay off in the long run. Thus it becomes an urgent issue to make rule-based systems available as open-source so that the development of new systems can be focused on creating the language resources.

1. Introduction

In the early days of machine translation, simplistic binary dictionaries were the only language resources that were used, and the results were poor. An increasing understanding of the importance of strategies for word sense disambiguation and for including morphological as well as syntactic knowledge emerged. A rule-based paradigm emerged, realized as direct translation with ad-hoc translation rules or transfer-based translation based on full syntactic sentence structures. Well-known shortcomings with the first strategy were due to difficulties in covering all contexts that were to be handled by translation rules (e.g. SYSTRAN), and with the transfer-based approach in covering the sentence structure in all its variation (see e.g. Hutchins, J. and Somers, H. 1992). A problem that was shared for all translation systems was developing full-covering dictionaries, and adapting them to specific domains hereby reducing the number of alternative interpretations. Initially, dictionaries were basically handcrafted, in particular as regards the definition of translation relations. A major step forward was taken with the development of strategies for aligning parallel text, bi-text, sentence-wise, word-wise, and phrase-wise. These strategies formed the basis for automatically extracting translation relations for dictionary-building purposes, and so-called statistical translation (Brown, P. F. et al. 1993), or example-based translation (see e.g. Way, A. and N. Gough. 2005). Translation was based on aligned parallel corpora and language models of the target language. Using the alignment strategies for building translation dictionaries for rule-based systems promoted the quality of these

systems substantially. Another strategy that was introduced to overcome problems with insufficient coverage of source language analysis grammar in transfer-based systems was using partial analyses (see e.g. Weijnitz et al. forthcoming.). Two basic strategies, statistical mt, and rule-based MT making use of partial analyses and the corpus-based translation dictionaries, emerged. Further, methods for automatically measuring similarity of the machine translated text with a reference translation, hereby estimating the quality of the machine translation text, were presented and heavily made use of in statistical as well as rule-based translation.

Each strategy has its shortcomings. A major problem with statistical MT concerns the identification of the bad translations, reasons behind them and ways of correcting errors of certain types, such as wrong or omitted lexical information, and syntactic errors. In contrast to rule-based systems, there are no individual linguistic rules to improve, rather global measures such as extending the language model or the translation model with more data, fine-tuning statistical parameters, and including external dictionaries. In other words, there are few linguistic ways of improving the translation. In rule-based translation, on the other hand, rules can be added and refined. Not surprisingly, current research aims at including linguistic knowledge into the statistical systems, and statistical models into rule-based systems. Ways of combining the two strategies into hybrid systems are explored (Hearne, M. 2005).

2. Outline

Here we will consider a situation where MT for a new language requires the building of a new system rather than

extending an existing (commercial) system with a new language pair. The situation appears e.g. for minority languages and less used languages or languages that for other reasons are not considered to be commercially motivated. Based on the availability of language resources and tools, a decision has to be made between a rule-based, a statistical or an example-based approach. For an illustration of the kinds of issues that may have to be considered in making the choice, we will present our experience from working in the rule-based and the statistical paradigm with MT of Swedish and English. As regards the example-based paradigm it is, basically, outside of our experience and will not be further discussed in this paper. First we will outline the development of a rule-based system with fall-back strategies, Convertus[1], and then the achievements made with statistical MT from and to Swedish. In working with different domains and typologically different languages, we made some observations concerning translation quality in the different experimental settings. Typically, a setting is characterized by parameters such as language pair, translation direction, and domain. The domain is defined by a corpus, and we found it motivated to take into account, not only the size of the corpus, but also features concerning repetition and overlap of text segments. Based on these criteria we introduce the term corpus *density* for capturing the openness/closeness of the domain. We will get back to these issues and their implications for translation quality in terms of BLEU. We will then present an experiment of a rule-based and a statistical approach to the translation of Swedish into English, and discuss the pro's and cons' of the two approaches. Finally, we draw some conclusions that seem to have some general interest.

3. Building a transfer-based system for translation of Swedish

As a result of more than ten years of research and development we arrived at a system for translating Swedish into English with a transfer-based core and complementary strategies for handling data outside the language description. It generates satisfactory results in several domains (automotive literature, agriculture, education). In the procedure leading to this system, the following main phases may be distinguished:

- Designing, implementing, and testing a modular, unification-based core engine, Multra (Beskow, B. 1993, Sågvall Hein 1994) with dictionaries and grammars of experimental size for translation of Swedish into English, German, and Russian; grammars and dictionaries were hand-crafted.
- Scaling up the system for one domain, e.g. automotive service literature, and one translation pair, Swedish to English; the scaling-up process was a major under-taking turning the prototype into a real system, Mats (Sågvall Hein, A. et al. 2003) capable of processing real-world documents. For the scaling-up effort a corpus of 16.1k sentence pairs (50,000 tokens) was established for training and testing, the so-called Mats corpus. Much effort was devoted to scaling up the dictionaries making use of word alignment techniques (Tiedemann, J. 2003), and to

organizing the lexical material in a database with built-in morphology. The modularity of the core engine is reflected in the database, which includes a source dictionary, a target dictionary and a translation dictionary in terms of translations relations between lexical units. Options for evaluating the coverage of the language description (dictionaries, grammars, transfer rules) and tracing the processing at a detailed level were also built into the system (Sågvall Hein et al. 2003). It should be mentioned, however, that the goal of scaling up of the grammars to cover the training corpus was not achieved in the project.

- Compensating for gaps in the language description by adding techniques for using partial analyses, external dictionaries, statistical models and fall-back strategies (Weijnitz et al. forthc); building an evaluation center to support systematic evaluation of translation quality (Forsbom, E. 2003); this phase lead to a major reorganization of the architecture and the process control, and motivated a new name of the system, i.e. Convertus.
- Progressively training the system for several domains to a BLEU score on the training text of about 0.5 on an average; user feed-back in the various projects in which the training took place indicates that a BLEU score of 0.4-0.5 based on a single reference corpus, is good enough for post-editing.
- Work in process concerns the automatic extraction of grammar from corpora (Megyesi 2002, Nivre 2005) for experimenting with different parsers and different languages, hereby making the system easily adaptable to new language pairs.

In conclusion, a well functioning translation system for translating Swedish into English in several domains was developed with a substantial investment of man power during many years. The system as such is not limited to translation from Swedish to English, but so far there are no language resources of relevant size for other language pairs. Achievements of general interest include techniques for building monolingual and bilingual dictionaries from parallel text as well as a database technology for storing and maintaining lexical data with inbuilt morphology (Tiedemann 2002). Another language-independent achievement is a flexible architecture permitting the plugging-in of modules for analysis, transfer, and generation, for on-line consultation of external dictionaries, and for using fall-back strategies for recovering processing in different problematic situations, e.g. when grammars are insufficient. However, for high-quality translation, grammars are required, for analysis, as well as transfer and generation. This may be a bottleneck in applying the technology to new languages, in particular, less used languages. With the further development of machine-learning techniques for extracting grammar from text, this problem may be reduced. Several modules of Convertus might be presented as open-source software. However, before that, the modules need to be properly packaged and documented.

An alternative strategy when it comes to approaching new languages may be to use statistical machine translation. Thus, now let us turn to our experience of applying statistical machine translation to Swedish.

---

4. Experiments with statistical machine translation of Swedish

To get a grasp of the perspectives of statistical machine translation from and to Swedish, we carried out a number of experiments. They indicate that language differences and translation direction have an impact on the translation quality measured by BLEU, in addition to corpus size and corpus density. The same system, to be described below, was used for all experiments.

4.1. The system

Phrase based systems work with both words and phrases, using at least two knowledge sources. Both the translation model and the language model are usually obtained automatically from parallel corpora. By using either the source-channel model, the more general direct maximum entropy translation model, or some other method, the translation model and language model are combined (Och, F. J. and Ney H. 2002).

Pharaoh is a beam search decoder implementing the best-performing methods for statistical machine translation as of year 2004 (Koehn 2004). The translation models were created using UPlug (Tiedemann, J. 2003), GIZA++ (Och, F. J. and Ney, H. 2000) and Thot (Ortiz-Martínez, D. et al. 2005). We used a basic set of models; a 3-gram target language model and a phrase translation model P(target|source), and a length penalty parameter. The length and model parameters were automatically optimized on development corpora. For our experiments, we restricted the source phrase lengths to 4. The language models were created using SRILM (Stolcke, A. 2002).

4.2. Language differences and translation direction

The first experiments were run on the Mats corpus. It includes source documents in Swedish with translations into several languages, among them English and German. Swedish as well as English, belong to the Germanic language family. However, English is felt to be closer, and easier to learn for a Swede than German.

The system was trained on 15.8k sentence pairs per language, sv-en, and, sv-de, and for each language pair 300 sentence pairs were kept aside and used for testing. As expected, English to Swedish outperforms English to German in a significant way (table 1).

Language pair	BLEU
sv->en	0.627
en->sv	0.646
sv->de	0.491
de->sv	0.506

Table 1: BLEU scores for the Mats corpus

Reversing the translation direction, translating from Swedish to English and German, respectively, implies a slight decrease in the BLEU value (table 1), i.e. from 0.646 to 0.627 for English and from 0.506 to 0.491 for German (cf. Papineni, K. A. et al. 2002). In other words, reversing the translation direction does not seem to have any real importance. Still the data will be further examined, in particular from a linguistic point of view.

We also made an experiment with Swedish -> Turkish using a sub-domain (information about Sweden) of a Swedish-Turkish parallel corpus (Megyesi et al., LREC 06) for training. Swedish is the source language. All in all, the sub-domain comprises 1289 sentence pair and the test corpus 206. Evidently, the training corpus is too small for successful SMT, and, in addition, the domain is fairly open. Further, Turkish belonging to the Altaic language family, is typologically very different from Swedish, being a Germanic language in the Indo-European language family. All things taken together, we cannot expect much, and in table 2 we present the results in both directions. Here we observe that the results are slightly better for Swedish as the source language than as the target language, as opposed to the case with English and German. Still the difference is very small, and hardly statistically significant. We need more data to investigate this aspect.

Language pair	BLEU
sv->tr	0.170
Tr->sv	0.156

Table 2: BLEU scores for the Turkish corpus

The experimental findings presented above inspired us to proceed in investigating the implications of the concrete language pair with regard to the quality of SMT. Thus we made a number of experiments for Swedish and other languages using Europarl (Koehn, P. 2005)

BLEU from sv	BLEU to sv	lang. pair	size: sent.
0.2403	0.2090	sv-es	20893
0.2065	0.2074	sv-es	10630
0.1238	0.1401	sv-es	1601
0.2382	0.2192	sv-pt	20726
0.2218	0.2099	sv-pt	10663
0.1288	0.1249	sv-pt	1601
0.1750	0.2194	sv-nl	20690
0.1592	0.1969	sv-nl	10645
0.1020	0.1471	sv-nl	1602
0.1814	0.1910	sv-fi	20663
0.1670	0.1708	sv-fi	10632
0.1048	0.0874	sv-fi	1601

Table 3: BLEU scores for Europarl

We used three different corpus sizes for training, i.e. in terms of sentences pairs: approximately, 20k, 10k and 1,6k. (Cf. the Turkish experiment with a training corpus of 1,2k sentence pairs.) As expected, as the corpus size grows, so does the BLEU score. We may also observe, that the best results are achieved for Swedish – Spanish, closely followed by Swedish- Portuguese. For both these languages, BLEU scores higher for Swedish as a source language than as a target language. Spanish and

Portuguese are Romance languages, and as might be expected, behave in a similar way in relation to Swedish. As regards Dutch, a German language, and Finnish, a Finno-Ugric language, the results are not conclusive. With Swedish as the source language, BLEU scores slightly higher for Finnish, but with Swedish as the target, Dutch outperforms Finnish fairly well, 0.21 versus 0.19. We would have expected a higher score for Dutch being a Germanic language, and as such closer to Swedish; here the statistical, corpus-based data contradict the typological tradition. The data will be further investigated. It is fair to assume, that Swedish-English and Swedish-German should keep their positions as best and second best in the score ranking list, as evidenced by the experiments on the automotive corpus.[2] They were run on a corpus of comparable size, i.e. ~16k sentence pairs. However, a source of error may be due to the openness/closeness of the domain. Europarl is assumed to represent a more open domain than the Scania corpus, and in accordance with earlier research (see e.g. Weijnitz et al. 2004) we expect these aspects to have an impact on the translation results. We will get back to corpus size and density in 3.3. below.

4.2.1. Measuring language differences

In view of the impact of language differences on statistical translation, we would like to have access to a similarity measure for judging the potential of statistical machine translation when approaching a new language pair. In theoretical linguistics language differences are investigated in the sub-field of typology. Here focus is set on identifying distinguishing features such as word order, grammatical relations, case markings, animacy, etc. These features can hardly be translated into a formal, computable measure. What seems to be required is a corpus-based measure that can be calculated automatically.

An option that comes into mind is to base such a measure on word alignment scores. For Swedish – English and Swedish – German, F-values of word alignment scores, based on a gold standard, have been presented (Tiedemann 2003). The author reports an F-value of 83.0 for Swedish – English, and 79.3 for Swedish – German. The figures are derived from the Mats corpus, i.e. the same corpus as was used for training in the SMT experiment presented above. The difference in F-value is comparable to that of BLEU for the two languages, i.e. a BLEU score of 0.65/0.63 (depending on translation direction) for English versus an F-value of 83.0, and a BLEU score of 0.51/0.49 for German versus an F-value of 79.3. The idea of using word alignment scores as a basis for measuring language differences with a view on statistical MT may seem like a circular reasoning; the word aligment phase is crucial in SMT and constitutes a major part in implementing such a system. Still the idea seems worth exploring further. We may consider building a matrix of F-values of word alignment scores for a large variety of languages to be consulted when considering the potential of SMT for a specific translation task. It seems preferable to base this inventory of language similarity measures on properly balanced corpora. As SMT performance depends on training corpus alignment quality, it is important that all language pairs involved are equally well aligned. For a start Europarl (Koehn, P. 2005) or JRC-Acquis (Steinberger, R. 2006) could be useful. In addition, gold standards have to be provided, or other means for calculating the global success of the word alignment process.

4.3. Corpus size and density

Our hypothesis is that there are more training corpus features than size that influence SMT performance. By corpus density we mean to what extent the corpus is repetitive at the sentence and n-gram levels. Table 4 shows the characteristics of a set of corpora, and their BLEU scores obtained when used for SMT. Percents show the type/occurrence ratios of the sentences and n-grams. In this test, there is a negative correlation between the BLEU scores and the sentence ratios. There is also a negative correlation, albeit weaker, between BLEU and n-gram ratios, and larger n-grams mean stronger correlation. The measures relate to the source side of the parallel corpus, only. As for the alignment quality, a crucial issue in SMT, we have no separate data. Still we assume that density correlates not only with the BLEU score but also with alignment quality as such. This, however, remains to be demonstrated. We conclude that corpus density, in addition to size, is a useful criterion when judging the prospect of SMT based on a parallel corpus.

5. Rule-based and statistical translation

There is a huge difference in effort between creating the language resources for a rule-based translation system and for building a statistical translation system. On the other hand, an SMT critically depends on access to a large parallel corpus of high quality within a fairly closed domain. The required size of the corpus cannot be estimated in the general case. It depends on the language pair to be translated, the translation quality to be achieved, and the density of the corpus. Above we have given some clues that should be useful in this context. A high-quality parallel corpus is almost also a sine non qua in building the language resources for a rule-based system, and in developing and evaluating the system. Spending efforts in building such a corpus should pay off in the final end, regardless of what strategy is chosen.

5.1. An experiment with Convertus and Pharaoh

For a comparison between rule-based and statistical translation, we made an experiment using Convertus for the rule-based approach and Pharaoh, as above, for the statistical approach. The experiment was run on the Scania 98 corpus, and the translation direction was Swedish to English.

System	BLEU	Pair	Corpus
convertus	0.377	sv->en	Scania 98
pharaoh	0.324	sv->en	Scania 98

Table 5: BLEU for Convertus and Pharaoh

---

[2] Unfortunately, data on SMT for Swedish-English and Swedish-German with regard to Europarl were unavailable at the time of writing due to technical problems.

BLEU from sv	BLEU to sv	Corpus	Pair	Sentences	Sentence unique	1-gram unique	2-gram unique	3-gram unique	4-gram unique
0.6274	0.6460	mats	sv-en	16716	60.9%	9.1%	44.8%	68.2%	78.2%
0.5201	0.5267	plug	sv-en	22195	61.3%	10.0%	50.8%	79.2%	89.1%
0.4653	0.4710	agri	sv-en	40910	79.6%	6.6%	38.9%	68.6%	80.1%
0.3523	0.3270	plugfgord	sv-en	4997	99.6%	12.5%	52.5%	83.7%	95.2%
0.2951	0.2944	plugjoinall	sv-en	9204	99.8%	12.1%	53.0%	85.1%	96.1%
0.2297	0.2654	plugfbell	sv-en	4207	100.0%	17.2%	63.9%	91.8%	98.6%

Table 4: Corpus density and BLEU

The training of Pharaoh as well as Convertus was based on the Mats corpus, and there was a token overlap between training data and test data of 5.8%, and a type overlap of 5.3%. This is a small overlap to be compared to the token overlap of 31.7%, and type overlap of 29.9% in the experiment on the Mats korpus (Table 1). As expected, there is a substantial difference in the BLEU score, i.e. 0.324 (Scania 98) and 0.627 (Mats corpus). As for Convertus, the training of the language resources was, basically, limited to the vocabulary; thus there is a big potential for further training of the grammars for this text type (analysis, transfer, as well as generation). In spite of that, Convertus scores higher than Pharaoh, even though the difference is fairly modest.

5.1.1. Translation quality

As has been shown before (e.g. Weijnitz et al. 2004), there is a correlation between the BLEU score and human assessment of translation quality. Further, according to our previous experience, a BLEU score below 0.4 - 0.5 is not good enough for post-editing. An informal study of the two versions of the machine-translated text confirms this view.

Most of the errors produced by both systems are due to words unknown the systems. The standard action taken by them both is to leave the word un-translated. As a fall back strategy, Converts consults an external dictionary outside the domain. This sometimes leads to a wrong choice. However, Convertus optionally provides a list of unknown words including both words that were left untranslated and words for which an external translation was chosen. This list provides the basis for up-dating the dictionary. Unknown words seem to cause worse problems in the SMT system, since there are several cases where the unknown word is simply missing. There is no "place-holder" of the problematic word, which makes the error hard to trace and the translation sometimes incomprehensible. Typically, the statistical system has problems with the translation of Swedish compounds, due to the difficulties of recognizing multiword units in the word aligment phase. An example may be *lufttorkarens function* [air dryer operation, the function of the air dryer] translated as *air function*. Evidently, only the first part of the compound was recognized by the word aligner and the error came out as a case of missing word, and as such hard to identify. More often, however, Swedish compounds are left untranslated e.g. *regulatorfjädrarna* [governor springs]. As mentioned above, the training of Convertus with regard to the Mats corpus was, basically, limited to the vocabulary. Shortcomings concerning the transfer

grammar appear in the translation of phrasal expressions, in particular phrasal verbs. Both systems expose structural errors and errors in the inflection of verbs and nouns. Convertus mainly encounters these problems when the subject is not located and when the noun is ambiguous in number. For the SMT system the distribution of these errors is less predictable.

The most striking difference between the two systems is the amount of omissions produced by the SMT. We didn't calculate their number in this experiment, but in a previous study (Weijnitz et al. 2004) the SMT system exposes more than four times as many instances of omission as the rule-based system. There may be differences in the settings of the two experiments, but the general impression seems to hold. In general, the cause of the errors produced by Convertus can easily be traced and explained in linguistic terms. The SMT system on the other hand, often produces less predictable errors, such as *svåra personskador* [serious personal injury] translated as *svåra not followed*. There is no obvious way to trace the cause of these errors to parameters in the translation or language model.

The rule-based system achieved better results but there is much room for improvements in both systems and a lot of common problems to be solved. Convertus can be improved by adjusting the grammars and extending the dictionaries. The statistical tools require larger amounts of domain-specific training data for better coverage and higher translation quality.

6. Conclusions

We outlined the development of a robust, rule-based MT-system for Swedish, Convertus, equipped with techniques for using partial analyses, external dictionaries, statistical models and fall-back strategies. Most modules of the system are candidates for open-source software, but before that they have to be properly packaged and documented. The system has been progressively trained for several domains to a BLEU score on training data of ~0.5%. User feed-back indicate, that a BLEU score of 0.4-0.5% generated by Convertus represents a translation quality good enough for post-editing.

We also made several experiment withs SMT for Swedish to find out more about factors influencing the translation quality. The experiments showed that language differences are an important issue, with BLEU scores ranging from 0.175 to 0.240 on the same corpus (Europarl). Translation direction turned out to be of minor importance. Corpus size and density were also

investigated, and a correlation was confirmed, not only between size and BLEU but also between density and BLEU. These factors should turn out to be useful in estimating the success of an SMT for a certain language pair and domain. However, our main conclusion will be that the investment made in developing a rule-based system, preferably backed up by a statistical system, will pay off in the long run. Thus it becomes an urgent issue to make rule-based systems available as open-source so that the development of new systems can be focused on creating the language resources. Regardless of strategy, the careful preparation of a parallel corpus appears as a crucial first step towards a high-quality MT system, where it is not readily available.

7. References

Beskow, B. (1993). Unification-Based Transfer in Machine Translation. In *RUUL 24*. Uppsala University. Department of Linguistics.

Brown, P. F. and Pietra, V. J. D. and Pietra, S. A. D. and Mercer R. L. (1993). The mathematics of statistical machine translation: parameter estimation. In *Computational Linguistics 19 2*. pp. 263-311. MIT Press, Cambridge, MA, USA

Forsbom, E. (2003). Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation. In *Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation*, MT Summit IX, pp. 29-36. New Orleans, Louisiana, USA, 27 September.

Hearne, M. (2005). Data-Oriented Models of Parsing and Translation. PhD Thesis, Dublin City University, Dublin, Ireland.

Hutchins, W. J. and Somers, H. L (1992) *An Introduction to Machine Translation*, London: Academic Press. ISBN 012362830X

Koehn. P (2005). Europarl: a parallel corpus for statistical machine translation. In *Tenth Machine Translation Summit, AAMT*. Phuket, Thailand. November 2005.

Koehn. P (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, In *AMTA 2004,* pp. 115-124.

Megyesi, B. (2002). Data-Driven Syntactic Analysis - Methods and Applications for Swedish. Ph.D.Thesis. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Nivre, J. (2005) Inductive Dependency Parsing of Natural Language Text. PhD Thesis, Växjö University.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models, In *ACL00* pp. 440—447. Hongkong, China, October 2000.

Och, F. J. and Ney H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL 2002* pp. 295-302.

Ortiz-Martínez, D. and García-Varea, I. and Casacuberta , F. (2005). Thot: a Toolkit To Train Phrase-based Statistical Translation Models. In *Tenth Machine Translation Summit, AAMT*. Phuket, Thailand. November 2005.

Papineni, K.A. and Roukos, S. and Ward, T. and Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL.*

Steinberger, R., and Pouliquen B. and Widiger, A. and Ignat, C. and E. Toma and Tufiş, D. and Varga, D. (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver

Sågvall Hein, A., (1994) Preferences and Linguistic Choices in the Multra Machine Translation System. In: Eklund, R. (ed.), NODALIDA '93 *Proceedings of '9:e Nordiska Datalingvistikdagarna',* Stockholm 3-5 June 1993.

Sågvall Hein, A. and Forsbom, E. and Tiedemann, J. and Weijnitz, P. and Almqvist, I. and Olsson, L.-J. and Thaning, S. (2002). Scaling up an MT Prototype for Industrial Use - Databases and Data Flow. I Proceedings from the Third International Conference on Language Resources and Evaluation (LREC'02), pp. 1759-1766, Las Palmas de Gran Canaria, Spanien, 29-31 maj.

Sågvall Hein, A and Weijnitz, P, and Forsbom, E, and Tiedemann, J. and Gustavii, E. (2003). MATS - A Glass Box Machine Translation System. In *Proceedings of the Ninth Machine Translation Summit*, pp. 491 – 493. New Orleans, USA, September 23-27, 2003.

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

Tiedemann, J. (2002). MatsLex - a multilingual lexical database for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volume VI. Las Palmas de Gran Canaria, Spain.

Way, A. and N. Gough. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering* 11(3):295--309.

Weijnitz, P. and Sågvall Hein, A. and Forsbom, E. and Gustavii, E. and Pettersson, E. and Tiedemann. J. (forthcoming). The machine translation system MATS - past, present & future. In *Proceedings of RASMAT'04. Uppsala, Sweden, 22-23 April.*

Weijnitz, P. and Forsbom, E. and Gustavii, E. and Pettersson, E. and Tiedemann, J. (2004). MT goes farming: Comparing two machine translation approaches on a new domain. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),* volym VI, s. 2043-2046. Lissabon, Portugal, May 26-28.