Expanding Parallel Resources for Medium-Density Languages for Free

Georgi Iliev, Angel Genov

Department of Computational Linguistics, Institute for Bulgarian, Bulgarian Academy of Sciences
52 Shipchenski prohod, bl. 17, Sofia 1113, Bulgaria
georgi@dcl.bas.bg, angel@dcl.bas.bg

Abstract

We discuss a previously proposed method for augmenting parallel corpora of limited size for the purposes of machine translation through monolingual paraphrasing of the source language. We develop a three-stage shallow paraphrasing procedure to be applied to the Swedish-Bulgarian language pair for which limited parallel resources exist. The source language exhibits specifics not typical of high-density languages already studied in a similar setting. Paraphrases of a highly productive type of compound nouns in Swedish are generated by a corpus-based technique. Certain Swedish noun-phrase types are paraphrased using basic heuristics. Further we introduce noun-phrase morphological variations for better wordform coverage. We evaluate the performance of a phrase-based statistical machine translation system trained on a baseline parallel corpus and on three stages of artificial enlargement of the source-language training data. Paraphrasing is shown to have no effect on performance for the Swedish-English translation task. We show a small, yet consistent, increase in the BLEU score of Swedish-Bulgarian translations of larger token spans on the first enlargement stage. A small improvement in the overall BLEU score of Swedish-Bulgarian translation is achieved on the second enlargement stage. We find that both improvements justify further research into the method for the Swedish-Bulgarian translation task.

Keywords: parallel corpora, automatic paraphrasing, compound splitting

1. Background

Monolingual paraphrasing has been shown to improve the performance of statistical machine translation (SMT) systems trained on parallel corpora of limited size which would otherwise suffer from sparse data (Callison-Burch et al., 2006). It is a way to augment existing parallel resources "for free", i.e. without the need to collect more parallel data, but by way of introducing variations in the source language which are considered identical to the original meaning. A number of experiments along these lines are described in the literature involving high-density languages, such as English and Spanish (Callison-Burch et al., 2006), (Nakov, 2008). We on the other hand choose to study its practical application in the case of a language pair for which no large parallel texts are available. Such resource scarcity precludes the use of some of the techniques described earlier, such as extracting paraphrases through a pivot language (Callison-Burch et al., 2006). Instead we resort to corpus-based techniques for compound analysis and linguistically motivated techniques for shallow parsing, morphological analysis and generation, using freely available monolingual tools and resources.

2. Goals

Swedish and Bulgarian are official languages of the EU and are both represented in the Europarl (Koehn, 2005) corpus. Bulgarian, a South-Slavic language using the Cyrillic alphabet, was added only recently, and the size of the English-Bulgarian version of the Europarl is rather modest – 226,768 sentence pairs compared to 1,678,333 sentence pairs in the Swedish-English Europarl. There are other Swedish-Bulgarian parallel texts available, such as the ones contained in the OPUS (Tiedemann et al., 2004), yet they are either literary texts of limited size, film subtitles of varying quality, or highly specialized texts. Apart from the Europarl and the JRC-ACQUIS, there are no other known available parallel language resources on general topics for the Swedish-Bulgarian language pair of any significant size.

Through intersection of the Swedish-English and the Bulgarian-English Europarl we arrive at a Swedish-Bulgarian parallel corpus of a total of 160,000 sentences to work with. Our primary goal is to enlarge the Swedish-Bulgarian text by combining several paraphrasing methods for the source language and to investigate the feasibility of the methods in the case of the languages at hand. In doing so we follow a shallow strategy, aimed at achieving as much coverage as possible by investing as little linguistic knowledge as possible. Our secondary goal is to study the effect from such enlargement on the quality of the output of an SMT system trained on the enlarged corpora compared to the results from training on the original corpus using standard metrics.

3. Language specifics

3.1. Compounding

Compounding is highly productive in Swedish (Sjöbergh and Kann, 2004) and accounts for a great number of unknowns in an unseen text, even if the separate parts of the compound are common. A prevailing number of compounds in a given Swedish text are endocentric noun-noun compounds. In terms of meaning most of them are determinative in that the first member of the compound modifies in some way the second member (Liljestrand, 1993). Typically, they would be translated into Bulgarian by a paraphrase where the first member is an adjective modifying the second member, as in the case of **järnrör** → желязна тръба (*iron*$_{\textbf{adj}}$ *pipe*), or by a paraphrase where the two members of the original compound appear in inverse order and are connected syntactically by a function

word, as in the case of **järnrör** → тръба от желязо (*pipe of iron*). Both translations pose a challenge to SMT, the first in terms of selecting the adjectival interpretation of the first member, and the second in terms of reordering and selecting the required function word. By splitting source compounds into their underlying components and further by paraphrasing the splits syntactically we expect to be able to reduce the number of unknowns and to render the source text into a sequence closer to that of the target language.

3.2. Rich morphology

Being inflected for gender and definiteness, Swedish nominals have a richer paradigm than English. This implies a higher number of unknowns on the source-language side in SMT. In order to be able to produce a translation of each token in the input, the SMT system needs to be trained on a parallel corpus containing the exact token. In a parallel corpus of limited size we cannot expect for each wordform in the source language to occur, and a system trained on such parallel corpus will normally either return empty output in the place where the translation of the unknown token is supposed to be, or return the input token unchanged in the final translation (Callison-Burch et al., 2006). One way to reduce such output in the case of morphologically complex languages would be to ensure that as many wordforms as possible occur in the training corpus. By introducing such morphological variations using automated means we expect to be able to reduce the number of unknowns in the source text even further.

4. Techniques

4.1. Compound splitting

For the purpose of compound splitting we chose to apply the method described by (Koehn and Knight, 2003) for splitting of German compounds, modified to reflect the specifics of Swedish noun-noun compounding. According to this method candidate compounds are split into the most likely sequence of tokens occurring in a monolingual corpus. Likelihood is expressed by a score combining the frequency counts of potential members of the candidate compound from the corpus, with a penalty for a larger frequency count of the candidate compound than the frequency count of any of its potential members.

In order to avoid overgeneration we limit our method to two-member candidate compounds analyzed as one of two main formal categories of determinative compounds, stem-form[1] compounds (*stamkomposita*) and case-form[2] compounds (*kasuskomposita*) (Liljestrand, 1993). In stem-form compounds the first member is the unmarked form of the underlying word or its stem. In case-form compounds the first member is a form of the underlying word marked for case.

4.2. Paraphrasing

4.2.1. Compounds

Analytic paraphrases of compounds are generated from compound splits by introducing function words and by generating a possessive expression. We proceed on the assumption that any such compound *ab*, where *a* is the first member and *b* is the second member, can be paraphrased as either a noun phrase in the form *b x a*, where *b* is the head and *x a* is a prepositional phrase modifying the head, or as $a_{\textbf{definite:possessive}}$ *b*. *x* is a function word from a list of the most typical function words that appear in this position. The list of function words is created on the basis of observations on a development set. The definite possessive form of the first member is extracted from the Swedish wordform lexicon SALDO (Borin et al., 2008).

A list of candidate paraphrases is generated for each compound split and each candidate paraphrase is passed as search term in a query to a search engine on the web[3]. Using the web as corpus we select the candidate paraphrase that occurs most frequently and discard all other candidates from the list. No compound paraphrase is introduced if none of the candidates on the list returns a search result.

4.2.2. Noun heads modified by a prepositional phrase

Noun phrases where the head is modified by a prepositional phrase are extracted from the original source language by means of shallow parsing using the SPARK (Megyesi, 2002) shallow parser for Swedish. For the purpose of paraphrasing we chose to apply the method described by (Nakov, 2008) for syntactic paraphrases of English. It is conservative and is expected to preserve the meaning of the original while adding syntactic variation. The English example "players of the Swedish team" can be paraphrased by the described method as:

(1) Swedish team players

(2) Swedish team's players

The premodification mechanism of Swedish morphology, however, is different. Compounding is a particularly productive premodification model, where adjectives and nouns, among others, can be compounded with other nouns, in theory without limitation. Thus "spelare i [det svenska] landslaget" (*players of the [Swedish] team*) can be rendered as:

(3) [svenska] landslag**s**spelare

(4) [svenska] landslaget**s** spelare

We proceed on the assumption that any observed structure in the shallow parse of the source language in the form **[NP* N_1 *NP] [PP* Prep [NP* N_2 *] *PP]**[4], where *Prep* is a preposition from a list of the most frequent prepositions that can appear in this position, and N_1 and N_2 are nouns, can be paraphrased as either a compound of type *ab* where *a* is N_2 and *b* is N_1, or as $N_{2\textbf{possessive}}$ N_1. The preposition list is created on the basis of observations on a development set.

[1] Author's English translation.

[2] See footnote 1 above.

[3] Bing, www.bing.com.

[4] SPARK parenthesis representation.

A considerable linguistic and computational effort is required to tackle the issue of Swedish compound generation. Formulating (3) above from an observed analytic sequence in the original corpus requires knowledge of the specific interfix that needs to be used when compounding the Swedish "landslag" (itself a compound) and "spelare". In the absence of a ready solution for corpus-based compound generation we apply an oversimplified model of combining candidate compound members by either the **-s-** interfix, which is by far the most frequent (Liljestrand, 1993), or without an interfix whatsoever in the cases where the first member ends in a vowel.
A morphological analysis is also required in order to be able to formulate a compound structure. Compare

(5) spelare i landslaget$_{\textbf{definite}}$ and

(6) landslag$_{\textbf{indefinite}}$**-s-**spelare

The base forms needed to fill the first position in the compound frame in (6) are extracted from the Swedish SALDO wordform lexicon.

4.2.3. Possessive structures

The SPARK shallow parser for Swedish does not split possessive expressions into two noun phrases allowing us to reverse the method already applied to creating analytic paraphrases of compound structures, so that compounds can be generated out of possessive expressions. We proceed on the assumption that any observed shallow structure in the form **[NP* N$_{1\textbf{definite:possessive}}$ N$_2$ *NP]** can be paraphrased as a compound of the type *ab*$_{\textbf{definite}}$ where *a* is N_1 and *b* is N_2. For example in

(7) **[NP* landsbygdens/NCUSG@DS utveckling/NCUSN@IS *NP]**

the (shallow) NP "landsbydgens utveckling" (*the rural areas' development*) is analyzed without breaking up, where the first member ("landsbygdens", *the rural areas'*) is in the definite, possessive form, and the second member ("utveckling", *development)* is in the indefinite. An automatic paraphrase using a compound made up of the two members is "landsbygdsutvecklingen" which results in 37,100 Bing results.

4.2.4. Limitations

It is obvious that the methods described so far are oversimplified in certain respects and could result in nonsensical output, in particular in terms of compound generation and possessive paraphrasing. This, however, does not run contrary to the goals of our experiment. The intuition behind this is that such nonsensical structures will not appear in the test set in the source language created by native speakers of Swedish, hence they will be filtered out from the phrase table used in actual translation.

4.3. Increasing wordform coverage

For the purpose of increasing the coverage of the corpus to include unobserved wordforms of lemmas which occur in a given form in the source language of the parallel corpus we resort to the language resources of the Grammatical Framework (Ranta, 2011) for Swedish. Consider, for example, the following chunk from the source language in the development set:

(8) **[PP* Enligt/SPS [NP* en/DI@US@S uppskattning/NCUSN@IS *NP] *PP]**

In this particular case we are able to generate all possible morphological variants of the Swedish noun phrase "en uppskattning" (*an evaluation*) using readily available resources from the Grammatical Framework for Swedish without violating the grammaticality of the original sentence. The corresponding noun phrase in the target language does not necessarily have to be in the same morphological form (singular, indefinite) as can be shown from the following example from the development set:

(9) Source: Alla säger samma sak: lösningen är att fortsätta att minska vårt beroende av (en enda energikälla) ***singular*:indefinite**

(10) Bulgarian target: Всички казват едно и също: че решението е да продължим да намаляваме нашата зависимост от (единични енергийни източници) ***plural*:indefinite**

(11) English target: *Everyone is saying the same thing: that the solution is to continue to reduce our dependency on single sources of energy*

The underlying intuition is that since the target-language noun phrase is not restricted to the morphological form of the source-language noun phrase, different morphological forms of noun phrases on the source-language side can be treated as paraphrases corresponding to one and the same target noun phrase.
In the case referred to in (8) we generate 3 additional sentence pairs where the original noun phrase "en uppskattning" is replaced by the following:

(12) uppskattningen (*the evaluation*)

(13) uppskattningar (*evaluations*)

(14) uppskattningarna (*the evaluations*)

We cannot, however, apply this approach indiscriminately to any noun phrase and preserve grammaticality because of the need for subject-verb agreement in many cases. In order to ensure that agreement rules are not violated we limit our transformations to noun phrases that are complements to a prepositional head in a prepositional phrase. The annotation needed to isolate those cases was already provided by the tools used for annotating the Swedish. This limitation cannot filter out all possible agreement errors but it is in line with the shallow strategy adopted in this work notwithstanding.

5. Results

5.1. Parallel data

The tools used in this work were developed on a development set of 17,663 sentences extracted from the Swedish-Bulgarian intersection of the Europarl corpus. A fresh

training set of 46,843 sentences ("Baseline") was extracted and used for training of the baseline Swedish-Bulgarian and Swedish-English SMT systems. The described paraphrasing methods were applied on the same training set resulting in three enlarged versions of the original training set which were used for training of the experimental Swedish-Bulgarian and Swedish-English SMT systems. The first enlarged version of 109,855 sentences ("Stage I") was produced using the methods described in 4.1 and 4.2.1 for compound splitting and compound paraphrasing. The noun-phrase variations discussed in 4.2.2 and 4.2.3 were introduced in the second enlarged version of 117,637 sentences ("Stage II"). The third enlarged corpus of 189,953 sentences ("Stage III") contains also the morphological variations discussed in 4.3.

Parallel corpora for the Swedish-Bulgarian and the Swedish-English translation task	Sentences
Baseline	46,843
Stage I	109,855
Stage II	117,637
Stage III	189,953

Table 1: Training corpora used for evaluation

5.2. Error analysis

We carried out a manual revision of two samples from the enlarged corpora of 300 sentences each in order to produce an analysis of the errors of the paraphrasing methods applied on Stage I and Stage II respectively.

5.2.1. Compound paraphrasing using the web as corpus

The first 300-sentence sample is the result of Swedish compound splitting and paraphrasing verified using the web as corpus. Paraphrases were checked by hand and failed (F) if deviating from the underlying analysis of the candidate compounds as determinative noun-noun compounds, and passed (P) otherwise. Due to the regularity of occurrence and the preserved grammaticality, adjectival compounds modified by a noun were also passed. Table 2 contains a summary of the typical candidates that were erroneously analyzed as determinative noun-noun compounds.

Of the 300-sentence sample 208 paraphrases were found to be correct according to the above criteria. Some of the errors identified can readily be eliminated, such as the ones involving closed-class words and highly productive prefixes and suffixes which were not filtered at this stage.

5.2.2. Generation of compounds and possessive expressions using heuristics

The second 300-sentence sample is the result of possessive paraphrasing of Swedish noun phrases where the head is modified by a prepositional phrase, and the compounding of Swedish possessive expressions. Table 3 summarizes the types of errors identified as a result of a manual revision.

Of the 300-sentence sample 180 paraphrases were found to be correct. The larger part of the identified errors were either due to improper shallow parsing or due to improper compounding or improper usage of definite/indefinite forms. At this stage these errors can readily be reduced by web searches, as indicated by the results from the evaluation of the first 300-sentence sample.

Paraphrasing errors	F/P
Verbal compounds, e.g. "godkänner" → "känner till god"	F
ab compounds where *a* is verb, e.g. "sittplats" → "plats för sitt"	F
ab compounds where *a* is adjective, e.g. "snabbtåg" → "tåget med snabb"	F
Nonsensical splits, e.g. "fordringar" → "ringar på ford"	F
Proper semantic rewrites, incorrect Swedish grammar, e.g. "arbetsmarknadssituationen" → "arbetsmarknadens situationen"	**P**
Incorrect function words, e.g. "problemområden" → "områden om problem"	F
Meaningful, but superfluous splits, e.g. "femtedelar" → "femte delar"	F
Splits of closed-class words or resulting in a closed-class member, e.g. "fastän" → "än med fast"	F
Splits where a member coincides with a highly productive prefix or suffix, e.g. "vanära" → "ära till van"	F
Accidental proper rewrites, e.g. "rättstvister" → "tvister om rätt"	F
Accidental proper rewrites of compound adjectives, e.g. "könsrelaterad" → "relaterad till kön"	**P**
Rewrites with a shift of meaning, e.g. "ostasien" → "asien till ost"	F

Table 2: Compound paraphrasing errors

5.3. Testing

Testing was performed on the Moses (Koehn et al., 2007) SMT system. We tested a total of four translation systems on each of the Swedish-Bulgarian and the Swedish-English translation tasks. The same Bulgarian and English language models built from the 160,000 sentences from the Swedish-Bulgarian intersection of the Europarl corpus were used in all test runs. No tuning was performed and the same uniform distortion and translation-model weights as well as the same language-model- and word-penalty weights were used to make sure that both the baseline and the experimental systems run under the same (suboptimal) conditions. A fresh test set of 10,000 unseen sentences was extracted from the Swedish-Bulgarian intersection of the Europarl corpus.

All three stages of enlargement resulted in an increase in source-language word coverage in the phrase tables extracted from the parallel texts, as shown in Table 4. The small differences between the number of source-language types used for training for the Swedish-Bulgarian and the English-Bulgarian translation task, respectively, result from filtering technicalities prior to training. Otherwise

Paraphrasing errors
Errors from improper shallow parsing, e.g. “frågan om vad” → “vads fråga”
Stem-form compounding errors, e.g. “jordbrukarnas intressen” → “jordbrukareintressena”
Improper definiteness/indefiniteness, “graden av överensstämmelse” → “överensstämmelses grad”
Possessive form of words that are not grammatically inflected, e.g. “konsekvenserna av detta” → “dettas konsekvenser”
Possessive rewrites with a shift of meaning, e.g. “Informationen om Europeiska rådet” → “Europeiska rådets information”
Case-form compounding errors and the **-s-** interfix, e.g. “konsumenternas intressen” → “konsumentsintressena”
Possessive rewrites of fossilized/idiomatic expressions, e.g. “effektivitetens namn” → “effektivitetsnamnet”
Possessive rewrites where compounding would be more suitable, e.g. “systemet med utresevisum” → “utresevisums system”

Table 3: Possessive paraphrasing errors

exactly the same source-language training data is used in both translation tasks on all stages.

Phrase table	Swedish-Bulgarian		Swedish-English	
	Swedish words	Bulgarian words	Swedish words	English words
Baseline	40,461	40,222	40,348	20,722
Stage I	41,239		41,125	
Stage II	44,403		44,281	
Stage III	47,628		47,495	

Table 4: Types occurring in phrase tables

5.4. SMT evaluation

Evaluation of the machine translation task is based on the standard BLEU metric (Papineni et al., 2002) as calculated by the *mteval* scoring script for the NIST Open Machine Translation 2009 Evaluation[5]. Table 5 shows the BLEU scores for the four tests on each translation task.

The Swedish-Bulgarian translation system trained on Stage I received the same overall BLEU score as the baseline. On this stage there was a small but consistent increase in the BLEU score of the translations of larger token spans as shown in Table 6. The Swedish-Bulgarian translation system trained on Stage II showed a slight improvement over the baseline. The Swedish-Bulgarian translation system trained on Stage III showed a slight deterioration.

None of the Swedish-English translation systems trained on any of the enlarged corpora showed an improvement over the baseline.

Corpus	BLEU scores	
	Swedish-Bulgarian	Swedish-English
Baseline	24.38	**35.87**
Stage I	24.38	35.81
Stage II	**24.44**	35.82
Stage III	24.34	35.57

Table 5: Overall BLEU scores

[5]www.itl.nist.gov/iad/mig/tools/

Baseline individual n-gram scoring								
1	2	3	4	5	6	7	8	9
56.25	**30.73**	**19.02**	11.95	7.65	5.04	3.36	2.27	1.58
Baseline cumulative n-gram scoring								
1	2	3	4	5	6	7	8	9
54.77	**40.48**	**31.19**	**24.38**	19.23	15.32	12.29	9.92	8.06
Stage I individual n-gram scoring								
1	2	3	4	5	6	7	8	9
56.16	30.67	19.01	**11.97**	**7.68**	**5.08**	**3.38**	**2.30**	**1.60**
Stage I cumulative n-gram scoring								
1	2	3	4	5	6	7	8	9
54.73	40.44	31.18	**24.38**	**19.26**	**15.35**	**12.32**	**9.96**	**8.10**

Table 6: BLEU individual and cumulative n-gram scoring of the Swedish-Bulgarian translation task

6. Discussion of results

Our work was inspired by the successful results reported in earlier studies of SMT systems for high-density languages, such as Spanish and English. Few, if any, results are reported on SMT between „smaller“ languages. Therefore we see this work as a feasibility study of the proposed methods, and the results as preliminary.

The manual revision of the two paraphrasing samples shows that there are areas that can be immediately improved in order to refine the automatic paraphrases and to reduce the noise introduced in the enlarged corpora by the paraphrasing methods we develop for Swedish, thereby improving the overall quality of the training corpora.

The enlarged corpora used for training the experimental SMT systems are produced as a result of the cascaded application, among others, of one data-driven tool (the Hunpos tagger (Halácsy et al., 2007) with a Swedish language model (Brants, 2000)) and two rule-based tools (the SPARK shallow parser and the Grammatical Framework). This naturally introduces a certain amount of noise, which we believe can have a negative impact on SMT quality.

The negative development in performance in the Swedish-English translation task can be attributed to factors specific to the language pair at hand. The paraphrasing methods used on the source-language side were to a large extent developed in view of the specifics of Bulgarian as target language (such as the analytic rewrites of compound words). This underscores the importance of the proposed methods for the Swedish-Bulgarian translation task.

The BLEU score is calculated on the basis of the number of exact n-gram matches between the machine translation and a reference translation produced by a human translator, where a BLEU score closer to 1 or 100% indicates higher similarity to the reference translation. As such it is found by some authors to be insensitive to the type of changes introduced by paraphrasing (Callison-Burch et al., 2006). Furthermore we note a considerable difference between the number of types in the Bulgarian and the English training data (40,222 vs. 20,722, Table 3), which most likely reflects the difference in morphological complexity between the two languages. This creates a bias in BLEU scoring in favour of English due to the considerably smaller search space, which is demonstrated by the difference in performance already on the baseline. It can be seen as a movement from a higher-dimensional (morphologically-rich) to a lower dimensional (morphologically-poor) space, where some loss of meaning and nuance is harmless (Lopez, 2008). Therefore we have reasons to believe that the gains from the paraphrasing methods we develop towards improving performance on Bulgarian as target language are higher than indicated by the BLEU score alone.

7. Conclusion and future work

We show a consistent improvement in the BLEU scores of translations of 4- and larger n-grams on Stage I and an overall improvement in the BLEU score on Stage II of the Swedish-Bulgarian translation task. Given the deficiencies identified in the course of evaluation of the tested paraphrasing methods this result is a success in itself and a strong reason to believe that more improvement can be achieved by refining the proposed paraphrasing methods.

The comparative evaluation of the results from the Swedish-Bulgarian and the Swedish-English translation tasks conducted under uniform conditions of the source-language training data indicates certain inherent challenges to the Swedish-Bulgarian machine translation task which have not been addressed previously and which we successfully approach by language-dependent monolingual paraphrasing.

The error analysis shows that in the future we should work towards the development of a robust corpus-driven method for compound generation and that all automatic paraphrases should be verified against monolingual data. The slight drop in performance on Stage III most likely indicates that morphological variations should be introduced more restrictively in the training data.

8. Acknowledgments

The research work described in the paper was partially supported by the Project Multilingual Parallel Corpora – in Aid of Contemporary Language Technologies (MU03), financed by the Program "Young Researchers" of the Bulgarian National Science Fund.

9. References

Lars Borin, Markus Forsberg, and Lennart Lönngren, 2008. *SALDO 1.0 (Svenskt associationslexikon version 2)*. Språkbanken, Göteborgs universitet.

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *In Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics Companion Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*, pages 79–86.

Birger Liljestrand. 1993. *Så bildas orden: handbok i ordbildning*. Studentlitteratur.

Adam Lopez. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3):8:1–8:49, August.

Beata Megyesi. 2002. *Data-Driven Syntactic Analysis - Methods and Applications for Swedish.* Ph.d.thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Preslav Nakov. 2008. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 338–342, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. Center for the Study of Language and Information/SRI.

Jonas Sjöbergh and Viggo Kann. 2004. Finding the correct interpretation of swedish compounds, a statistical approach. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 899–902.

Jörg Tiedemann, Lars Nygaard, and Tekstlaboratoriet Hf. 2004. The opus corpus – parallel and free. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1183–1186.