Statistical Machine Translation with Readability Constraints

Sara Stymne, Jörg Tiedemann, Christian Hardmeier and Joakim Nivre

Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden

firstname.lastname@lingfil.uu.se

ABSTRACT

This paper presents experiments with document-level machine translation with readability constraints. We describe the task of producing simplified translations from a given source with the aim to optimize machine translation for specific target users such as language learners. In our approach, we introduce global features that are known to affect readability into a document-level SMT decoding framework. We show that the decoder is capable of incorporating those features and that we can influence the readability of the output as measured by common metrics. This study presents the first attempt of jointly performing machine translation and text simplification, which is demonstrated through the case of translating parliamentary texts from English to Swedish.

KEYWORDS: Machine Translation, Text Simplification, Readability.

1 Introduction

Typically, statistical machine translation (SMT) focuses on the translation of isolated sentences. However, humans usually emphasize the translation of coherent texts into equally coherent translations targeted at a specific audience. In this paper, we address the problem of including document-wide features into statistical MT that may influence the style of the generated documents in the target language. For this, we adopt the task of producing simplified translations that could be useful, for example, for language learners, dyslectic people or simply for non-experts who want to grasp the major content in highly domain-specific documents written in a foreign language. An example for the latter could be legal texts that often use a very domain-specific terminology and jargon.

Readability and text simplification has been widely studied in the field of computational linguistics and several metrics and approaches have been proposed in the literature. Common readability metrics make use of global text properties such as type/token ratios, lexical consistency, and the proportion of long versus short words as indicators. Our goal is to incorporate these features in machine translation in order to combine text simplification and adequate translation in one system. To the best of our knowledge, this has not been attempted before and represents a novel and challenging idea in the field of MT research.

Global features such as the ones mentioned above require new approaches to the general problem of decoding in SMT. Fortunately, we have recently presented a new document-level decoder, which, contrary to standard SMT decoders, translates documents as a unit instead of sentences in isolation (Hardmeier et al., 2012). This allows us to define document-wide features in the target language to test our ideas. Our application is also a good test case for the capabilities of the decoder and we would like to use our findings in future developments of general user-targeted machine translation.

The contributions of this paper are thus two-fold: (1) We show that document-wide decoding can effectively use global features and (2) we demonstrate that readability features can be used in SMT to produce simplified text translations. The remainder of the paper is organized as follows: First, we introduce important background on document-level decoding and readability. Thereafter, we present our experiments using a set of global features. Finally, we add some information about related work, summarize our findings and give ideas about future work.

2 Document-wide SMT

Most current SMT systems translate sentences individually, assuming independence between the sentences in a text. This independence assumption is exploited in the most popular SMT decoding algorithms, which efficiently explore a very large search space by using dynamic programming (Och et al., 2001). Integrating discourse-wide information into traditional SMT decoders is difficult because of these dynamic programming assumptions. We therefore implement our document-level readability models in the recently published document-level SMT decoder Docent (Hardmeier et al., 2012), which does not have these limitations.

The model implemented by Docent is phrase-based SMT (Koehn et al., 2003). The decoder uses a local search approach whose state consists of a complete translation of an entire document at any time. The initial state is improved by applying a series of operations using a hill climbing strategy to find a (local) maximum of the score function. The three operations used are to change the translation of phrases, to swap the position of two phrases , and to resegment phrases. This setup is not limited by dynamic programming constraints, so we can define

scoring functions over the entire document. By initializing the decoder state with the output of a Moses run, which makes it possible to include all models except the ones with document-level dependencies, we ensure that the final hypothesis is no worse than what would have been found by Moses alone.

3 Readability – Metrics and Features

Readability is a complex notion that is related both to properties of texts, and to individual readers and their skills. Chall (1958) defines four elements that she considers significant for a readability criterion: vocabulary load, sentence structure, idea density, and human interest. Mühlenbock and Kokkinakis (2009) map these categories to metrics for readability that can be used in combination in order to capture several readability aspects. Vocabulary load can be measured as word-frequency ratios, or based on the proportion or number of long words. For Swedish, they suggest the commonly used metric *LIX* (Björnsson, 1968), which measures sentence length and proportion of long words and in addition the proportion of *extra long words* (XLW). For sentence structure, parsed text can be used to calculate the proportion of simple versus complex sentences. Such calculations are relatively costly, however, and dependent on the availability of parsers and definitions of sentence types, and they suggest that *average sentence length* (ASL) can be used as a proxy. Idea density can be measured based on lexical variation and the proportion of function vs content words. A common density measure for Swedish is the *Word variation index* (OVIX), a reformulation of *type token ratio* (TTR) that is less sensitive to text length. Another one is *nominal ratio* (NR), which is a ratio based on parts of speech. Finally, human interest can be captured for instance by the proportion of proper names (PN). To cover the four readability aspects, we use all these metrics for evaluation. Their definitions are shown in Eq. 1–6, where $C(x)$ is the count of x.

$$\text{ASL} = \frac{C(\text{tokens})}{C(\text{sentences})} \tag{1}$$

$$\text{LIX} = \text{ASL} + 100 * \frac{C(\text{tokens} > 6 \text{ chars})}{C(\text{tokens})} \tag{2}$$

$$\text{XLW} = 100 * \frac{C(\text{tokens} >= 14 \text{ chars})}{C(\text{tokens})} \tag{3}$$

$$\text{OVIX} = \frac{\log(C(\text{tokens}))}{\log\left(2 - \frac{\log(C(\text{types}))}{\log(C(\text{tokens}))}\right)} \tag{4}$$

$$\text{NR} = \frac{C(\text{nouns}) + C(\text{prepositions}) + C(\text{participles})}{C(\text{pronouns}) + C(\text{adverbs}) + C(\text{verbs})} \tag{5}$$

$$\text{PN} = \frac{C(\text{proper names})}{C(\text{tokens})} \tag{6}$$

In order to couple text simplification with SMT we also need to include features that can capture some aspects of readability into the decoder. While our main focus is on document-level features, we also include some basic sentence level count features used in readability metrics. On the document-level we use features for type token ratio and for lexical consistency. Table 1 gives an overview of the readability features.

Document-level		Sentence-level	
OVIX	Word variation index (Eq. 4)	SL	Sentence length in words
TTR	Type token ratio (Eq. 7)	nLW	Number of long words (> 6 chars)
Qw	Q-value, word level (Eq. 8)	nXLW	Number of extra long words (>= 14 chars)
Qp	Q-value, phrase level (Eq. 8)		

Table 1: Readability features used in the decoder

Lexical consistency has not typically been used as a readability indicator as such, but consistent vocabulary can be considered likely to improve readability. To measure it we chose the Q-value metric, which has been proposed for measuring bilingual term quality (Deléger et al., 2006). It is based on the frequency and consistency in translation of term candidates, as shown in Eq. 8 where $f(st)$ is the frequency of the term pair and $n(s)$ and $n(t)$ are the numbers of different term pairs in which the source and target terms occur respectively. We include a Q-value feature both on word level and on phrase level. On the phrase level we consider the phrases used by the SMT decoder, and on the word level we consider individual source words, and their alignment to $0-N$ target words.

$$\text{TTR} = \frac{C(\text{tokens})}{C(\text{types})} \tag{7}$$

$$\text{Q-value} = \frac{f(st)}{n(s)+n(t)} \tag{8}$$

4 Experiments

In the following, we show results for our experiments with the Docent decoder that include readability features and compare them to runs without them. The systems are evaluated using both MT and readability metrics.

4.1 Experimental Setup

We evaluate our models on parliamentary texts from Europarl (Koehn, 2005), which contain both complex sentences and a lot of domain-specific terminology. All tests are performed for English–Swedish translation. Our system is trained on 1,488,322 sentences. For evaluation, we extracted 20 documents with a total of 690 sentences from a separate part of Europarl. A document is defined as a complete contiguous sequence of utterances of one speaker. We excluded documents that are shorter than 20 sentences and longer than 79 sentences.

Moses (Koehn et al., 2007) was used for training the translation model and SRILM (Stolcke, 2002) for training the language model. We initialized our experiments with a Moses model that uses standard features of a phrase-based system: a 5-gram language model, five translation model features, a distance-based reordering penalty, and a word counter. These features were optimized using minimum error-rate training (Och, 2003) and the same weights were then used in Docent. Currently, we are developing the optimization procedure in Docent and could not use it in this work. We thus used a grid search approach for choosing weights for the readability-based features with low, medium, and high impact relative to the standard features.

We performed automatic evaluations using a set of common metrics for MT quality and readability. For MT quality we used BLEU (Papineni et al., 2002) and NIST (Doddington, 2002),

Feature	Weight	BLEU↑	NIST↑	LIX↓	ASL↓	OVIX↓	XLW↓	NR↓	PN↑
Reference	–	–	–	50.47	24.65	57.73	3.08	1.055	0.013
Baseline	–	0.243	6.12	51.17	25.01	56.88	2.63	1.062	0.015
OVIX	low	0.243	6.11	51.00	25.09	54.65	2.60	1.069	0.015
	medium	0.228	5.83	49.33	25.45	44.43	2.53	1.063	0.015
	high	0.144	4.41	46.59	29.09	31.65	1.82	0.941	0.013
TTR	low	0.243	6.12	51.04	25.11	55.25	2.60	1.070	0.015
	medium	0.225	5.75	49.86	26.19	45.31	2.44	1.080	0.014
	high	0.150	4.48	48.30	30.54	32.95	1.77	0.975	0.012
Qw	low	0.242	6.10	51.16	25.07	57.16	2.62	1.064	0.015
	medium	0.231	5.90	51.28	25.32	58.90	2.62	1.074	0.015
	high	0.165	4.93	50.92	26.14	60.61	2.63	1.101	0.016
Qp	low	0.243	6.12	51.16	24.99	56.94	2.65	1.061	0.015
	medium	0.229	5.99	49.79	24.14	54.75	2.62	1.060	0.015
	high	0.097	3.90	41.45	21.99	39.22	2.39	1.129	0.015
nLW	low	0.244	6.14	50.96	24.98	56.73	2.63	1.065	0.015
	medium	0.225	5.96	46.72	24.21	55.39	2.72	1.080	0.018
	high	0.106	4.11	30.27	22.18	45.41	1.78	0.899	0.023
nXLW	low	0.241	6.10	51.03	24.96	56.69	1.85	1.060	0.015
	medium	0.225	5.85	50.92	25.09	56.56	0.19	1.070	0.016
	high	0.224	5.84	50.97	25.12	56.55	0.19	1.068	0.016
SL	low	0.242	6.21	51.07	24.22	57.79	2.71	1.058	0.016
	medium	0.211	5.94	50.77	21.61	60.93	3.15	1.040	0.018
	high	0.150	4.38	50.77	18.46	65.37	3.72	1.072	0.021

Table 2: Results of systems with single readability features, compared to the reference and baseline. Arrows indicate the direction of metrics (better MT/more readable). For definitions of metrics and features see Eq. 1–6 and Table 1.

and for readability we used LIX, ASL, OVIX, XLW, NR and PN, explained in section 3. Since we lack a customized evaluation set, the MT metrics were computed against a standard reference set of normal Europarl translations that are not simplified. We can thus expect that simplification leads to a decrease in MT metrics. To further investigate the effects on adequacy and readability, we performed a small human evaluation. We leave more principled optimization and evaluation to future work, where we also plan to use simple dev and test sets.

4.2 Results

In this section we present results on metrics and a small human evaluation. We also exemplify the types of simplifications we can achieve in our setup.

Metrics

Table 2 shows the results when we activate one readability feature at a time using low, medium, and high weights for each feature. We can see that the baseline and reference are quite similar with respect to readability with some interesting differences, for example for extra long words. As expected, giving a high weight to a readability feature usually results in a dramatic decrease in MT quality (with respect to the unsimplified reference translation), but also affects the corresponding readability feature(s) much, in some cases with unreasonably low scores (see e.g., LIX for the nLW feature). Using low or medium weights, on the other hand, can give reasonable MT scores as well as some improvements on several readability metrics. Obviously, the features corresponding directly to a metric affects that metric, such as LIX for nLW and OVIX for OVIX and TTR. Several features also affects other readability metrics, though. For

	BLEU↑	NIST↑	LIX↓	ASL↓	OVIX↓	XLW↓	NR↓	PN↑
Baseline	0.243	6.12	51.17	25.01	56.88	2.63	1.062	0.015
LIX (nLW+SL)	0.214	5.96	46.09	23.02	56.27	2.90	1.061	0.018
OVIX+SL	0.229	5.94	48.86	24.34	44.53	2.63	1.046	0.015
Qp+OVIX+nLW+SL	0.225	5.93	47.77	24.08	43.77	2.65	1.045	0.016
All features	0.235	6.04	49.29	24.34	47.80	1.98	1.046	0.015

Table 3: Results for systems with combinations of readability features (medium weights)

	Preferred system		
	Baseline	Equal	Readability (All)
Adequacy	51	33	16
Readability	33	29	38

Table 4: Preferred system with regard to adequacy and redability in the human evaluation

instance, OVIX and TTR decrease several metrics, but give an increase in sentence length, which is unwanted. For the Q-value, the effect is very different when used on phrase and word level. On the phrase level it decreases most metrics, except NR, which increases, while the effect is small on the readability metrics when used on the word level. We need to analyze these outcomes in more detail in the future.

In Table 3 we show results for some of the possible feature combinations, using medium weights. As expected, the effect on the readability metrics is more balanced in these cases. For the system with all features there are improvements on all readability metrics, except for PN, which is on par with the baseline. The other systems that use some global feature also have a positive effect on most readability metrics, while the LIX system that uses only local features has little effect on OVIX and a negative effect on extra long words. All these systems show only a modest decrease in MT quality, though. We can thus show that the decoder, with global features, managed to simplify translations on aspects corresponding to vocabulary load, idea density, and sentence structure while maintaining a reasonable translation quality.

Human Evaluation

We also performed a small human evaluation of 100 random non-identical sentences from the baseline and the system using all readability features.[1] For each sentence we ranked the output on adequacy, how well the content is translated, and readability, how easy to read the translations are. For annotation we used the Blast interface (Stymne, 2011), which also showed the overlap between the two translations. The evaluation was performed by the four authors, who are either native Swedish speakers, or have a good command of Swedish.

The results are shown in Table 4. It shows that the baseline produces a higher number of adequate translations than the system with readability features, but adequacy is also equal often. For readability there is a small advantage for the system with readability features, which is consistent with the improvement on readability metrics. Overall the output was often very similar, with only few words differing. In several of the cases where the baseline was judged as having better adequacy the cause is a single changed word, which can be more common or shorter, but has the wrong form or part of speech, which means it does not fit into the context. In other cases some non-essential information is removed from the sentence, which

[1] 177 out of 690 sentences were identical.

while making the translation less adequate, is actually what we want to achieve. There were several cases where essential words have disappeared from the translation as well, though.

This evaluation was small scale, with course-grained judgments, and for only one possible system with readability features. In the future we want to look at several systems in some more detail. We can still see a tendency though, that we can gain a little bit in readability, but we're currently paying a cost as concerns adequacy.

Translation Examples

In Table 5 we show sample translations, in order to exemplify the types of operations our current system is able to perform. One type of successful simplification is to remove words that are not crucial for the main content. In many of the systems with readability constraints, the phrase *the honourable Members* has been simplified, either by removing the adjective and giving only *ledamöterna* ('the members'), or even by using the pronoun *ni* ('you'). Another positive simplification is that of *in such a way that*, which is translated quite literally in the baseline, but simplified into *så att* ('so that') in several of the systems. There are also instances, however, where the changes lead to a loss of information, especially for the shorter translation options. Examples are *handlingsplan* ('action plan'), which is reduced to *plan* ('plan') in the nLW system, and *2003* which is missing in the OVIX and Qp systems. There are several cases where different translations have been chosen for a word or phrase. Sometimes this can lead to a simplification, as in the nLW system where the everyday expression *bli klar* ('finish') is used instead of the more formal *avsluta* ('finish'). In other cases the translation options are of a relatively similar degree of difficulty, such as *vissa/en del/några* ('some'). In some cases a change of translation also lead to a change of part of speech, as for *uppmärksamhet* ('attention') which is often translated as the adjective *uppmärksam* ('attentive'), which unfortunately have led to syntactic problems in these translations. In general, as can be expected of SMT, there are some problems with fluency in all translations, but they tend to get worse in the systems with high-weight readability features.

There are also other types of changes, which are not shown in Table 5. Using the feature for extra long words tend to break long compounds, sometimes successfully, for example, translating the long compound *gemenskapslagstiftningen* ('the community legislation') into the genitive construction *gemenskapens lagstiftning* ('the community's legislation'), which is done in the XLW and All systems. Sometimes this is less successful, however, e.g., when not translating *World Trade Organisation* at all or giving the English-based abbreviation *WTO* instead. There are also cases where the readability features lead to changes in syntactic structure, such as the translation of *the excellent work he has done* into *hans utmärkta arbete* ('his excellent work') in some systems with the Qw feature, instead of a literal translation in the baseline.

5 Related Work

As far as we are aware this is the first work presenting joint machine translation and text simplification. Aziz et al. (2012) investigate the task of translating subtitles where time and space constraints are important, which leads to the task of sentence compression, which is related to our work on simplifying translated texts. They introduce dynamic length penalties which they integrate in a standard SMT decoder. Their model successfully compresses subtitles on three data sets. However, they also show that a similar compression can be achieved with appropriate tuning data that meets the length constraints. There are also a number of studies that use SMT techniques for monolingual paraphrasing (e.g., Ganitkevitch et al., 2011) and

Source	As the honourable Members know - some speakers have mentioned it - the European Council at Lisbon paid particular attention to promoting our efforts to implement risk capital in such a way that the action plan will be finished in 2003.
Baseline	Som de ärade ledamöterna vet - vissa talare har nämnt det - som Europeiska rådet i Lissabon ägnat särskild uppmärksamhet åt att främja våra ansträngningar att genomföra riskkapital på ett sådant sätt att handlingsplanen kommer att vara avslutat år 2003.
All (medium)	Som ledamöterna vet - vissa talare har nämnt det - som Europeiska rådet i Lissabon särskilt uppmärksam på att främja våra insatser för att genomföra riskkapital så att handlingsplanen kommer att vara avslutat 2003.
LIX (medium)	Som ledamöterna vet - vissa talare har nämnt det - Europeiska rådet i Lissabon lagt särskild vikt vid att främja våra ansträngningar att genomföra riskkapital så att handlingsplanen kommer att vara avslutat år 2003.
OVIX+SL (medium)	Som ni vet - vissa talare har nämnt det - som Europeiska rådet i Lissabon särskilt uppmärksam på att främja våra ansträngningar att genomföra riskkapital så att handlingsplanen kommer att avslutas under 2003.
OVIX (high)	Som ledamöter - en del talare har nämnt det - som Europeiska rådet i Lissabon särskilt uppmärksam på att stödja våra insatser för att genomföra av riskkapital, på så sätt att handlingsplanen kommer att vara avslutat i.
Qp (high)	Som de ärade ledamöterna vet, som några talare har nämnt det rådet i Lissabon, ägnat särskild uppmärksamhet åt att vi för att genomföra riskerna i det att handlingsplanen kommer att avslutas med.
nLW (high)	Som ni vet - vissa har sagt det - EU:s möte i Lissabon lagt särskild vikt vid vår för att genomföra risk i så att den plan att bli klar under 2003.
SL (high)	Som ledamöterna vet vissa talare har nämnt - Europeiska rådet i Lissabon särskilt uppmärksammat främja våra ansträngningar att genomföra riskkapital så att handlingsplanen avslutas 2003.

Table 5: Examples of translation output from a sample of systems

sentence compression (e.g., Knight and Marcu, 2000; Specia, 2010). Furthermore, there is a wide range of publications using other methods for monolingual sentence compression and text simplification, (e.g., Daelemans et al., 2004; Cohn and Lapata, 2009).

Readability has also been investigated as an effect of text summarization, as measured by user studies (Margarido et al., 2008) and automatic metrics (Smith and Jönsson, 2011). In these studies the readability was generally better in the summarized texts than in the original texts. Stymne and Smith (2012) showed that SMT is affected by summarization, but found no relation between readability and SMT quality measured by standard evaluation metrics.

There is also related work concerned with the integration of wide contextual features in machine translation, such as lexical consistency. The effect of lexical consistency in translation has been studied by Carpuat (2009) and Carpuat and Simard (2012). Tiedemann (2010) proposed cached models to push consistent translation with some success in the case of domain adaptation. The use of word sense disambiguation in SMT is another example where wide contextual information can be incorporated on the source side (Carpuat and Wu, 2007; Chan et al., 2007)

Another study related to ours is Genzel et al. (2010), who study poetry translation and perform joint translation and poetry creation of news text as well as translation of poems that keep the

poetic form. They use features in the decoder such as rhyme and meter. They also introduce constraints over the target language output in order to adapt to the task-specific properties. However, they do not work on the document level, which would be an interesting direction for future work.

6 Conclusion and Future Work

In this article we explore a few readability related features in statistical machine translation. We have shown that these global features can successfully be integrated in a document-level decoder. We have evaluated a test case for English and Swedish using parliamentary texts which illustrates the effect of adding readability constraints. Our results demonstrate that the decoder can easily be influenced in terms of several aspects of readability of its output and that the approach can lead to a number of different types of simplifications. As expected, the translation quality goes down to some extent as measured by MT metrics as the one and only reference translation is not aimed at simplifying the text compared to its original version. The human evaluation also showed that we suffered on adequacy. So far we did not formally evaluate the effect on fluency, but the inspection of sentences showed that there were problems in this respect, which we plan to address in the future, for instance by the use of sequence models based on parts of speech or morphology.

Further directions for future work include the incorporation of additional features. So far, we only use surface features but we could complement them with features based on linguistic annotation such as POS labels and syntactic information. We could capture aspects measured by readability metrics such as NR and PN, or apply features like Q-value only for content words, but also could help to eliminate complex structures such as relative clauses. Another direction could be the task of splitting sentences into simpler ones if necessary. This however, would involve substantial developments in the decoder framework and would require appropriate training data that cover such cases. Finally, we are currently working on the optimization of feature weights within the document-level decoder. In our current experiments, no automatic tuning procedure has been applied for document-level features. We expect that proper weighting will be crucial to optimize the interaction between feature functions.

References

Aziz, W., de Sousa, S. C. M., and Specia, L. (2012). Cross-lingual sentence compression for subtitles. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 103–110, Trento, Italy.

Björnsson, C. H. (1968). *Läsbarhet*. Liber, Stockholm.

Carpuat, M. (2009). One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado.

Carpuat, M. and Simard, M. (2012). The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada.

Carpuat, M. and Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72, Prague, Czech Republic.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. Columbus : Bureau of Educational Research, Columbus, Ohio, USA.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 33–40, Prague, Czech Republic.

Cohn, T. and Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

Daelemans, W., Höthker, A., and Sang, E. T. K. (2004). Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1045–1048, Lisbon, Portugal.

Deléger, L., Merkel, M., and Zweigenbaum, P. (2006). Enriching medical terminologies: an approach based on aligned corpora. In *International Congress of the European Federation for Medical Informatics*, pages 747–752, Maastricht, The Netherlands.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Proceedings of the Second International Conference on Human Language Technology*, pages 228–231, San Diego, California, USA.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Durme, B. V. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1179, Edinburgh, Scotland.

Genzel, D., Uszkoreit, J., and Och, F. (2010). "Poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, Massachusetts, USA.

Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea.

Knight, K. and Marcu, D. (2000). Statistics-based summarization — Step one: Sentence compression. In *National Conference on Artificial Intelligence (AAAI)*, pages 703–710, Austin, Texas, USA.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the NAACL*, pages 48–54, Edmonton, Alberta, Canada.

Margarido, P., Pardo, T., Antonio, G., Fuentes, V., Aluísio, S., and Fortes, R. (2008). Automatic summarization for text simplification: Evaluating text understanding by poor readers. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana*, pages 310–315, Vila Velha, Brazil.

Mühlenbock, K. and Kokkinakis, S. J. (2009). LIX 68 revisited – an extended readability. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 160–167, Sapporo, Japan.

Och, F. J., Ueffing, N., and Ney, H. (2001). An efficient A* search algorithm for Statistical Machine Translation. In *Proceedings of the ACL 2001 Workshop on Data-Driven Machine Translation*, pages 55–62, Toulouse, France.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Smith, C. and Jönsson, A. (2011). Automatic summarization as means of simplifying texts, an evaluation for Swedish. In *Proceedings of the 18th Nordic Conference on Computational Linguistics (NODALIDA'11)*, Riga, Latvia.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, 9th International Conference (PROPOR'10)*, pages 30–39, Porto Alegre, Brazil.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA.

Stymne, S. (2011). Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies, Demonstration session*, Portland, Oregon, USA.

Stymne, S. and Smith, C. (2012). On the interplay between readability, summarization and MTranslatability. In *Proceedings of the 4th Swedish Language Technology Conference*, pages 71–72, Lund, Sweden.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the ACL 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP)*, pages 8–15, Uppsala, Sweden.