Towards improving English-Latvian translation: a system comparison and a new rescoring feature

Maxim Khalilov[*1], **José A. R. Fonollosa**[†], **Inguna Skadiņa**[‡], **Edgars Brālītis**[‡], **Lauma Pretkalniņa**[‡]

[*]Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam, The Netherlands
[†]Centre de Recerca TALP, Universitat Politècnica de Catalunya, Barcelona, Spain
[‡]Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia
m.khalilov@uva.nl, adrian@gps.tsc.upc.edu, inguna@latnet.lv, edgars.bralitis@gmail.com, lauma.pret@gmail.com

Abstract

This paper presents a comparative study of two alternative approaches to statistical machine translation (SMT) and their application to a task of English-to-Latvian translation. Furthermore, a novel feature intending to reflect the relatively free word order scheme of the Latvian language is proposed and successfully applied on the n-best list rescoring step. Moving beyond classical automatic scores of translation quality that are classically presented in MT research papers, we contribute presenting a manual error analysis of MT systems output that helps to shed light on advantages and disadvantages of the SMT systems under consideration.

1. Introduction

Translation into the languages with relatively free word order has received a lot less attention than translation into fixed word order languages (English), or into analytical languages (Chinese). At the same time this translation task is found among the most difficult challenges for machine translation (MT), and intuitively it seems that there is some space in improvement intending to reflect the free word order structure of the target language.

Non-configurational (free word order) languages differ crucially from the configurational languages (that follow a restrictive word order scheme), first of all, in the way how the pragmatic information is conveyed. In configurational languages (like, German, English, or Spanish) the order of syntactic constituents varies negligibly (or does not vary at all) and the emotional component of the message is usually transmitted through intonation variation[2].

In contrast to them, the non-configurational languages (like, Latvian, Russian, or Greek) often rely on the order of constituents to convey the topicalization or focus of the sentence. Latvian language is the target language in the experiments that we report in this paper. There are about 1.5 million native Latvian speakers around the world: 1.38 million are in Latvia, while others are spread in USA, Russia, Sweden, and some other countries. Also Latvian language is second language for about 0.5 million inhabitants of Latvia and several tens of thousands from neighbor countries, especially Lithuania[3].

Latvian is one of two living Baltic languages. It is charactarized by rich morphology, high level of morph-syntactic ambiguity and relatively complex pre- and postposition structures. Despite that it descends from the same ancestor language as Germanic languages, it differs from them significantly and the experience gained from machine translation into German or English can hardly be transferred to the English-to-Latvian translation task.

Not numerous attempts to model the free word order phenomena can be found in literature. For example, a thorough discussion of the appropriate word ordering strategy (using contextual information) for English-to-Turkish rule-based machine translation can be found in Hoffman (1996); in Zwarts and Dras (2007), the authors concentrate on SMT of indigenous Australian languages (one of the two languages under consideration is a prototypical non-configurational language).

Nowadays, scientific community is starting to express doubts that the models working pretty well for fixed word order languages are still efficient for free word order languages (for example, construction of an English-to-Czech SMT system taking into consideration very rich morphology and relatively free word order of Czech is one of the goals of the Euromatrix(plus) project[4]).

Despite the fact that translation from/to Latvian seems to be an extremely interesting task, this challenge has not received much attention in the SMT community. The first and only profound study on Latvian-to-English SMT, to our knowledge, was dated to 2008 (Skadiņa and Brālītis, 2008). Latvian was also one of the languages under consideration in a research conducted on all language pairs of the Aquis corpus Koehn et al. (2009).

In this paper, we study some aspects of English-to-Latvian MT. First, we compare the outputs of two SMT systems following two different approaches to MT and reporting results in terms of automatic evaluation metrics.

We consider a "de facto" standard phrase-based Moses system (Koehn et al., 2007) (factored and unfactored models) and an alternative N-gram-based SMT system (Mariño et al., 2006). We then study a *novel feature function* designed to reflect the non-configurational structure of the target lan-

[1]The bulk of the work presented in this paper was done during the first author's Ph.D studies in Centre de Recerca TALP, Universitat Politècnica de Catalunya, Barcelona (Spain).

[2]There are some exceptions to the general rule, for example, when it is necessary to emphasize the object of the sentence (*I agree **with you** -> **With you** I agree*), or in question sentences.

[3]Source: State Language Agency `http://www.valoda.lv/lv/latviesuval`

[4]`http://www.euromatrix.net/`

guage and show that the SMT can benefit from this feature introduced on the N-best-list rescoring/reranking step.
The paper concludes with human error analysis performed in order to identify the major strengths and weaknesses of the Moses and N-gram-based SMT systems when translating into Latvian.
The rest of this paper is organized as follows. Section 2 briefly describes phrase- and N-gram-based SMT systems, Section 3 introduces a new rescoring feature that reflects a non-configurational nature of Latvian language, Section 4 reports on the experimental setups along with automatic translation scores, in Section 5 we present results of human evaluation and error analysis, while Section 6 concludes the paper.

2. Two approaches to SMT

SMT is based on the principle of translating a source sentence ($f_1^J = f_1, f_2, ..., f_J$) into a sentence in the target language ($e_1^I = e_1, e_2, ..., e_I$). The problem is formulated in terms of source and target languages; it is defined according to equation (1) and can be reformulated as selecting a translation with the highest probability from a set of target sentences (2):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ p(e_1^I \mid f_1^J) \right\} = \quad (1)$$

$$= \arg\max_{e_1^I} \left\{ p(f_1^J \mid e_1^I) \cdot p(e_1^I) \right\} \quad (2)$$

where I and J represent the number of words in the target and source languages, respectively.
Modern state-of-the-art SMT systems operate with the bilingual units extracted from the parallel corpus based on word-to-word alignment. They are enhanced by the *maximum entropy approach* and the posterior probability is calculated as a *log-linear combination* of a set of feature functions (Och and Ney, 2002). Using this technique, the additional models are combined to determine the translation hypothesis $\hat{e}_1^I$ that maximizes a log-linear combination of these feature models (Brown et al., 1990), as shown in (3):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (3)$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.
There have been a bunch of publications that investigate the source of the possible improvements and degradations in translation quality when using translation systems underlined by different statistical models. For example, in Crego et al. (2005a), the N-gram-based system is contrasted with a state-of-the-art phrase-based framework, while in Khalilov and Fonollosa (2009), the authors seek to estimate the advantages, weakest points, and possible overlap between syntax-augmented MT and N-gram-based SMT. In Zollmann et al. (2008) the comparison of phrase-based, hierarchical, and syntax-based SMT systems is provided.
In this section we discuss the translation models compared in this work.

2.1. Phrase-based SMT

Most of modern state-of-the-art SMT systems follow the phrase-based approach to translation. The basic idea of this approach is to segment the given source word sequence into monolingual phrases, afterwards translate them and compose the target sentence (Och and Ney, 2002).
A phrase-based translation is considered as a three step algorithm: (1) the source sequence of words is segmented in phrases, (2) each phrase is translated into target language using translation table, (3) the target phrases are reordered to be inherent in the target language.
A bilingual phrase (which in the context of SMT do not necessarily coincide with their linguistic analogies) is any aligned pair of m source words and n target words that satisfies two basic constraints: (1) words are consecutive along both sides of the bilingual phrase and (2) no word on either side of the phrase is aligned to a word outside the phrase (Och and Ney, 2004). The probability of the phrases is estimated by relative frequencies of their appearance in the training corpus.
The system built for the English-to-Latvian translation experiments is implemented within the open-source Moses toolkit (Koehn et al., 2007). Standard training and weights tuning procedures which were used to build our system are explained in details on the Moses web page: `http://www.statmt.org/moses/`. Two word reordering methods are considered: a distance-based distortion model (see 2.1.1.) and lexicalized MSD block-oriented model (see 2.1.2.).

2.1.1. Distance-based

A simple distance-based reordering model default for Moses system is the first reordering technique under consideration. This model provides the decoder with a cost linear to the distance between words that should be reordered.

2.1.2. MSD

A lexicalized block-oriented data-driven MSD reordering model (Tillman, 2004) considers three different orientation types: monotone (M), swap (S), and discontinuous (D). MSD model conditions reordering probabilities on the word context of each phrase pair and considers decoding process a block sequence generation process with the possibility of swapping a pair of word blocks. Notice that in the experiments conducted within the framework of this study a MSD model was used together with a distance-based reordering model.

2.2. N-gram-based SMT system

Alternative approach to SMT is the N-gram-based approach (Mariño et al., 2006), which regards translation as a stochastic process that maximizes the joint probability $p(s, t)$, leading to a decomposition based on bilingual n-grams, typically implemented by means of a Finite-State Transducer (Casacuberta et al., 2002).
The core part of the system constructed in this way is a translation model (TM), which is based on bilingual units, called tuples, that are extracted from a word alignment according to certain constraints. A bilingual TM actually constitutes an n-gram LM of tuples, which approximates the

joint probability between the languages under consideration and can be seen here as a LM, where the language is composed of tuples.
The tuple-based approach is considered monotonous because the model is based on the sequential order of tuples during training. However, for a great number of translation tasks, a certain reordering strategy is required. In the framework of this study we consider a non-deterministic reordering method (see 2.2.2.).
As a consequence of a distinct representation of bilingual units, the N-gram-based approach differs from the $phrase$-based SMT by using a higher order HMM of the translation process. While regular phrase-based SMT considers context only for phrase reordering but not for translation, the N-gram-based approach conditions translation decisions on the previous translation decisions.

2.2.1. Additional features

The N-gram translation system implements a log-linear combination of five additional models:

- *an n-gram target LM;*
- *a target LM of Part-of-Speech (POS) tags*;
- *a word penalty model* that is used to compensate for the system's preference for short output sentences;
- *source-to-target and target-to-source lexicon models* as shown in Och and Ney (2004).

2.2.2. Extended word reordering

An extended monotone distortion model based on the automatically learned reordering rules was implemented as described in Crego and Mariño (2006). Based on the word-to-word alignment, tuples were extracted by an *unfolding* technique. It allows the generation of shorter tuples, increasing the system's reordering flexibility and, at the same time, alleviating the problem of embedded units. As a result, the tuples were broken into smaller tuples, and these were sequenced in the order of the target words (Crego et al., 2005b).
The reordering strategy is additionally supported by a 4-gram LM of reordered source POS tags. In training, POS tags are reordered according to the extracted reordering patterns and word-to-word links. The resulting sequence of source POS tags is used to train the n-gram LM.

2.2.3. Decoding and optimization

The open-source Marie[5] decoder was used as a search engine for the translation system. Details can be found in Crego et al. (2005b). The decoder implements a beam-search algorithm with pruning capabilities. All the additional feature models were taken into account during the decoding process. Given the development set and references, the log-linear combination of weights was adjusted using a *simplex* optimization method and an n-best re-ranking as described in `http://www.statmt.org/jhuws/`.

[5] `http://gps-tsc.upc.es/veu/soft/soft/marie/`

3. Repetition bonus feature

A word reordering in case of English-to-Latvian translation is a mapping between the words spanning the same semantic "spaces" but often located in different positions in English and Latvian. In the general case, there are multiple equivalent positions on the Latvian side.
Repetition bonus (RB) feature function is used on the n-best rescoring step and intends to reflect a non-configurational nature of the target language. The RB score RB_i^j shows how many times the word permutations (all combinations of words independently on their positions) in the translation hypothesis i appear in the list of best translations of the sentence j.
Formally, it is expressed as follows:

$$RB_i^j = \exp(a(i,j)) \tag{4}$$

where $a(i,j)$ is the number of times when η_i appears in N_j, N_j is a set of n-best translation of the sentence j, $1 \leq j \leq k$, $i \neq j$, k is the length of the translated corpus (measured in lines); $\eta_i \in P_i^j$, P_i^j is a set of possible word permutations in the translation hypothesis i of the sentence j. In other words, $a(i,j)$ is the number of times a permutation of the current translation hypothesis i of the sentence j.
Consequently, the n-best list enhanced with the RB feature is:

$$\begin{array}{lll}
1 \;|||\; \text{Hypothesis } 1 \;|||\; cost_1^{1_1} cost_2^{1_1} \ldots cost_{RB}^{1_1} \\
1 \;|||\; \text{Hypothesis } 2 \;|||\; cost_1^{1_2} cost_2^{1_2} \ldots cost_{RB}^{1_2} \\
\ldots \\
1 \;|||\; \text{Hypothesis } N_1 \;|||\; cost_1^{1_{N_1}} cost_2^{1_{N_1}} \ldots cost_{RB}^{1_{N_1}} \\
\ldots \\
j \;|||\; \text{Hypothesis } 1 \;|||\; cost_1^{j_1} cost_2^{j_1} \ldots cost_{RB}^{j_1} \\
j \;|||\; \text{Hypothesis } 2 \;|||\; cost_1^{j_2} cost_2^{j_2} \ldots cost_{RB}^{j_2} \\
\ldots \\
j \;|||\; \text{Hypothesis } N_j \;|||\; cost_1^{j_{N_j}} cost_2^{j_{N_j}} \ldots cost_{RB}^{j_{N_j}} \\
\ldots \\
k \;|||\; \text{Hypothesis } 1 \;|||\; cost_1^{k_1} cost_2^{k_1} \ldots cost_{RB}^{k_1} \\
k \;|||\; \text{Hypothesis } 2 \;|||\; cost_1^{k_2} cost_2^{k_2} \ldots cost_{RB}^{k_2} \\
\ldots \\
k \;|||\; \text{Hypothesis } N_k \;|||\; cost_1^{k_{N_k}} cost_2^{k_{N_k}} \ldots cost_{RB}^{k_{N_k}}
\end{array}$$

where $cost_a^{b_c}$ is the cost value for the feature function a calculated for the translation hypothesis c of the sentence b.
Less formal example can be found in Figure 1.

$$\begin{array}{lll}
j \;|||\; word_A word_B word_C \;|||\; cost_{j_1}^1 cost_{j_1}^2 \ldots 2 \\
j \;|||\; word_B word_A word_C \;|||\; cost_{j_2}^1 cost_{j_2}^2 \ldots 2 \\
j \;|||\; word_A word_B word_C word_D \;|||\; cost_{j_3}^1 cost_{j_3}^2 \ldots 1 \\
j \;|||\; word_B word_A word_C word_E \;|||\; cost_{j_4}^1 cost_{j_4}^2 \ldots 0 \\
j \;|||\; word_B word_A word_C \;|||\; cost_{j_5}^1 cost_{j_5}^2 \ldots 2 \\
j \;|||\; word_B word_C word_D \;|||\; cost_{j_6}^1 cost_{j_6}^2 \ldots 0 \\
j \;|||\; word_A word_B word_D word_C \;|||\; cost_{j_7}^1 cost_{j_7}^2 \ldots 1
\end{array}$$

Figure 1: *Example of the n-best list enhanced with the RB feature.*

The RB feature is a posterior probability on target-side word order and, to a certain extent, can be considered an

additional LM that aims to capture possible permutations of a particular set of words on the n-best list level.

3.1. Rescoring

A RB feature is integrated in the phrase- and N-gram-based SMT systems within a discriminative rescoring/reranking framework, which incorporates complex feature functions by using the entire translation hypothesis to generate a score.

During the first step, the decoder produces a list of n candidate translations based on the weights vector trained over the m basic features. Then, the statistical scores of each generated translation candidate are rescored using information provided by v additional features that presumably should add information not included during decoding to better distinguish between higher and lower quality translations. During this step, a rescoring vector is trained over $m+v$ features and provides different, better choices for the single-best translation hypothesis.

4. Experiments

We used JRC Acquis parallel corpus of about 5.4M running tokens in the Latvian part and of about 6.7M tokens in the English part of the corpus. Development set contained of 500 sentences randomly extracted from the bilingual corpus, test corpus size was 1000 lines. Both the datasets were provided with 1 reference translation. Main corpus statistics are shown in Table 1, including number of sentences, running words, vocabulary size and average sentence length.

	Latvian	English
Training		
Sentences	269.98 K	269.98 K
Words	5.40 M	6.65 M
Vocabulary	101.25 K	60.47 K
Development		
Sentences	0.50 K	0.50 K
Words	9.90 K	12.36 K
Vocabulary	3.08 K	2.30 K
Test		
Sentences	1.00 K	1.00 K
Words	20.18 K	24.64 K
Vocabulary	4.98 K	3.49 K

Table 1: Basic statistics of the English-Latvian JRC-Acquis corpus.

4.1. Experimental details

Phrase-based experiments were conducted following the guidelines provided on the Moses site (`www.statmt.org/moses/`). We used the 2008 version of Moses decoder. As an alternative to a traditional phrase-based model (*PB-u*), we considered a factored phrase-based SMT (*PB-f*) that constructed translation/generation models on the basis of the factorized corpus (preface words (word forms), POS tags, and lemmas for English and Latvian). Both configurations include MSD and distance-based reordering models commonly used in phrase-based SMT.

A detailed description of the N-gram-based system (*NB*), which was used in the work, can be found in Mariño et al. (2006). Notice that the monotone translation system was enhanced with a word graph input sentence representation providing the decoder with various reordering paths (Crego and Mariño, 2007).

Word alignments have been estimated using GIZA++ (Och and Ney, 2003) tool assuming 4 iterations of the IBM2 model, 5 HMM model iterations, 4 iterations of the IBM4 model, and 50 statistical word classes (found with mkcls tool (Och, 1999)).

A 4-gram target LM with unmodified Kneser-Ney backoff discounting was generated using the SRI language modeling toolkit (Stolcke, 2002) and was used in all the experiments.

Automatic evaluation was case insensitive and punctuation marks were not considered.

4.2. Results

Table 2 shows the results of translation, both starting with "standard" configurations, and contrasts them with the performance shown by the *RB-enhanced systems*. Best scores are placed in cells filled with grey (within phrase-based and N-gram-based experimental sets). We consider BLEU, NIST, PER, WER, and METEOR scores measured on the

System	Dev	Test				
		BLEU	NIST	PER	WER	METEOR
Phrase-based SMT (Moses)						
PB-u	42.69	43.95	78.91	38.48	51.47	59.85
PB-f	42.40	43.80	78.83	38.13	51.43	59.76
PB-u+RB	42.40	44.18	79.04	38.36	51.45	60.51
N-gram-based SMT (TALP)						
NB	43.52	45.11	82.40	35.05	47.98	62.52
NB+RB	43.59	44.86	82.10	35.23	48.22	62.54

Table 2: *English-to-Latvian experimental results.*

test dataset, along with the final point achieved as a result of the weight optimization procedure.
In the rescoring step, we calculate the same set of features that was used during decoding plus RB feature. All the rescorings were done on the basis of the 1000-best lists.

5. Human evaluation and error analysis

Human analysis of translation output allows going beyond automatic scores and, in the general case, provides a comprehensive comparison of multiple translation systems. On the first step, the *PB-u* and *NB* every non-repetitive test line was presented to the judge, who was instructed to decide that the two translations were of equal quality, or that one translation was better than the other. The results of the standard systems comparison can be found in Table 3.

	PB-u	NB	Equal
Preference	58	193	539

Table 3: Human evaluation results (standard systems).

Table 4 shows the results of preference analysis for the standard and RB-enhanced systems.
In addition, we performed error analysis on 100 first sentences from the test data. The analysis of typical errors generated by each system was done following the error classification scheme proposed in (Vilar et al., 2006) by contrasting the systems output with the reference translation. Table 5 presents the comparative statistics of errors generated by *PB-u* and *NB* baselines, as well as by RB-enhanced systems.
The number of missing content words in the output generated by the unfactored phrase-based system is more than five times higher than the analogous value for the N-gram-based system. We explain this difference by a high analytical inflection of the Baltic languages that is modeled better by the N-gram-based system since it involves surrounding context not only for phrase reordering, but conditions translation decisions on previous translation decisions.
Extra words embedded into the correctly translated phrases is another prominent source of errors generated by the phrase-based system (34, i.e. 13.6 % in case of *PB-u* and 9 (4.9 %) for the *NB* system). We explain it by the key difference in internal representation of translation units between phrase-based and N-gram-based SMT approaches. From the other hand, it illuminates the weakest point of the phrase-based systems having access to a small training material; the decoder relies only on a sparce set of phrases probabilities which does not provide an ideal path during beam search. On the contrary, N-gram-based SMT selects the partial translation hypothesis among a set of candidates based on the bilingual LM probabilities that find out to be more efficient for a given translation task.
However, the aforementioned feature of the N-gram-based architecture turns to be a weakness when dealing with (local) word reordering, that is reflected in the high number of reordering errors produced by the *NB* system. Experimental results show that internal phrase-based reordering enhanced with the MSD block-oriented model viewing translation as a monotone block sequence generation process outperforms the POS-based word graph reordering model used in N-gram-based experiments (22 local word/phrase order errors (8.8 %) coming from the *Pb-u* system vs. 37 errors of this type (20.3 %) produced by the *NB* system). At the same time, long-range word dependencies are modeled by *PB-u* and *NB* with comparable performance[6].
The difference in total number of errors is negligible, however a subjective evaluation of the systems output shows that the translation generated by the N-gram system is more understandable than the phrase-based one.

6. Discussion and conclusions

In this paper two alternative SMT systems are compared: the standard phrase-based and the N-gram-based SMT systems. The comparison shows that the N-gram-based SMT outperforms Moses-based translation system for the English-to-Latvian translation task in terms of automatic scores (≈1.3 BLEU points (3 %)) and human "best/worse" evaluation.
NIST and METEOR[7] correlate with BLEU results. In terms of PER and WER metrics, *NB* system outperforms *PB* configurations by about 3 points that can be interpreted as that the N-gram-based SMT can translate the context better and consistently produces less reordering errors than phrase-based system.
Translations generated by the N-gram-based system were preferred by the annotator more than 3 times more ofthen than the output of the phrase-based system.
Introduction of the novel feature intending to reflect the relatively free word order of the target language does not yield significant improvements in translation quality for the N-gram-based SMT, but allows improving results shown by the phrase-based system.
Human error analysis clarifies advantages and disadvantages of the systems under consideration and reveals the most important sources of errors for both systems.
Findings of this study, along with the robust error analysis of the SMT system outputs can be a very important step on the way of the translation quality improvement when dealing with free word order languages. A study on how the RB feature interacts with an n-gram LM and different smoothing methods can be an interesting research topic to be done in the future.
We have not addressed the inflectional aspect of Latvian and the associated data sparsness problem, that present many opportunities for future work on improving of English-to-Latvian translation.

Acknowledgments

Work was partially supported by the Spanish Ministerio de Educación y Ciencia (TIN2006-12767), by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project), and by the Latvian Council of Science

[6]For clarity's sake, it is important to notice that the English-to-Latvian translation task is not characterized by the big number of long-range reordering dependencies.

[7]Strictly speaking, METOR values can not be considered absolutely confident for Latvian due to the limited area of METEOR application (at present, it can be efficiently used for major languages only).

	PB-u	PB-u+RB	Equal	NB	NB+RB	Equal
Preference	27	353	728	79	74	637

Table 4: Human evaluation results (standard systems vs. RB-enhanced systems).

Type	Sub-type	PB-u	NB	PB-u + RB	NB + RB
Missing words		**64**	**16**	**68**	**31**
	Content words	52	10	54	9
	Filler words	12	6	14	22
Word order		**35**	**58**	**31**	**48**
	Local word order	11	23	8	16
	Local phrase order	11	14	10	12
	Long range word order	6	7	6	5
	Long range phrase order	7	14	7	15
Incorrect words		**128**	**82**	**132**	**84**
	Wrong lexical choice	25	20	23	27
	Incorrect disambiguation	10	4	11	9
	Incorrect form	51	46	52	35
	Extra words	34	9	38	9
	Style	8	2	8	3
	Idioms	0	1	0	1
Unknown words		**4**	**8**	**6**	**7**
Punctuation		**20**	**18**	**18**	**25**
Total		**250**	**182**	**245**	**185**

Table 5: Error statistics for a 100-line representative test set.

(project "Application of Factorized methods in English-Latvian Statistical Machine Translation System"). The authors want to thank Khalil Sima'an (Universiteit van Amsterdam) for his valuable discussions and suggestions.

7. References

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

F. Casacuberta, E. Vidal, and J. M. Vilar. 2002. Architectures for speech-to-speech translation using finite-state models. In *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, pages 39–44.

J. M. Crego and J. B. Mariño. 2007. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

J. M. Crego and J. B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

J. M. Crego, M. R. Costa-jussà, J. B. Mariño, and J. A. R. Fonollosa. 2005a. Ngram-based versus phrase-based statistical machine translation. In *Proc. of the 2nd Int. Workshop on Spoken Language Translation (IWSLT'05)*, pages 177–184.

J. M. Crego, J. B. Mariño, and A. de Gispert. 2005b. An ngram-based statistical machine translation decoder. In *Proceedings of INTERSPEECH05*.

B. Hoffman. 1996. Translating into free word order languages. In *Proceedings of COLING'96*, pages 556–561, Copenhagen, Denmark, August.

M. Khalilov and J. A. R. Fonollosa. 2009. N-gram-based statistical machine translation versus syntax augmented machine translation: comparison and system combination. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 424–432, Athens, Greece.

Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open-source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL) 2007*, pages 177–180.

Ph. Koehn, A. Birch, and R. Steinberger. 2009. 462 machine translation systems for europe. In *Proceedings of the twelfth Machine Translation Summit*, pages 65–72, Ottawa, Ontario, Canada, August.

J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.

F. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)02*, pages 295–302.

F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 3(4):417–449, December.

F. Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 71–76, June.

I. Skadiņa and E. Brālītis. 2008. Experimental statistical machine translation system for latvian. In *Proceedings of the 3rd Baltic Conference on HLT*, pages 281–286.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 901–904.

C. Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL'04*.

D. Vilar, J. Xu, L. F. D'Haro, and H. Ney. 2006. Error Analysis of Machine Translation Output. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702.

A. Zollmann, A. Venugopal, F. Och, and J. Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of Coling 2008*, pages 1145–1152, Manchester, August.

S. Zwarts and M. Dras. 2007. Statistical machine translation of australian aboriginal languages: Morphological analysis with languages of differing morphological richness. In *Proceedings of the Australasian Language Technology Workshop*, pages 134–142, Melbourne, Australia, December.