Pseudo-lemmatization in Croatian-English SMT

Marija Brkić, Maja Matetić
University of Rijeka
Department of Informatics
Radmile Matejčić 2, 51000 Rijeka
{mbrkic@inf.uniri.hr, majam@inf.uniri.hr}

Sanja Seljan
University of Zagreb, Faculty of Humanities and Social Sciences
Department of Information and Communication Sciences
Ivana Lučića 3, 10000 Zagreb
{sanja.seljan@ffzg.hr}

Abstract. *One of the first difficulties in conducting a thorough analysis of statistical machine translation involving Croatian as a morphologically rich and resource poor language is the lack of quality language resources. This paper presents results of two standard fourteen feature Croatian-English phrase-based statistical machine translation systems. Prior to building the second system a partial pseudo-lemmatization of the Croatian parts of training and test sets is made in an attempt to simplify the translation process. Besides automatic evaluation, a manual evaluation is conducted in order to gain insight into the nature of the translation differences achieved between the two systems.*

Keywords. phrase-based statistical machine translation, pseudolemmatization, Croatian-English

1 Introduction

This paper presents two newly built Phrase-Based Statistical Machine Translation (PBSMT) systems for translating from Croatian into English. This language pair has become particularly interesting since July 2013 when Croatia gained EU membership. One of the first difficulties in conducting a thorough experimentation with the standard PBSMT involving this language pair is the lack of quality language resources. Adequate resources, and therefore SMT systems, are mainly developed only for widespread languages. The domain of our choice is Acquis Communautaire domain since this domain, to the best of our knowledge, enables access to the biggest existent parallel corpus of the languages concerned. The practical result of this research is the creation of language resources necessary for SMT.

SMT systems are representatives of a data-driven approach to machine translation (MT) [5]. The data-driven approach to MT started its development in the 1980s, while the idea of SMT came out of the IBM research labs at the end of 1980s. Research on SMT has been facilitated by the tools developed by the participants of the John Hopkins University workshop which were made freely available, so SMT reached its full bloom around the end of the millennium [13]. The grounds of SMT are set out in [4]. In the context of the classification presented by Vaquois' triangle [10], PBSMT systems are one step above the direct approach, because they take word inflections into account. These systems are unidirectional, i.e. they can translate from one language into the other, but not vice versa. PBSMT systems [11, 18] are among the highest quality systems of nowadays [13]. The term phrase refers to a randomly selected sequence of words and not to its usual linguistic interpretation. An illustration of the translation process in the Croatian-English PBSMT systems is given in Fig. 1. Although the second sentence translation in the illustration is ill formed, it is provided to illustrate the reordering that occurs in the translation process.

The second section of the paper gives overview of research similar to the one presented here, with a special focus on Croatian, and south Slavic languages in general. At the beginning of the third section a concise contrastive analysis of Croatian and English is given. Since these languages have many differences, a morphological analysis of the Croatian parts of training and test sets is made prior to building the second system. The analysis is guided by the research in Goldwater and McClosky [9], Popović et al. [21, 22], Popović and Ney [23], and Maučec and Kačić [17]. The second part of the third section describes training and test sets used in the research. A detailed description of the systems built is given in the last part of the section three. Section four reports translation results of the two systems in terms of TER (Translation Edit Rate) [26], BLEU (Bilingual Evaluation Understudy) [19] and Meteor [1], and with regard to the system optimization procedure. Manual evaluation results

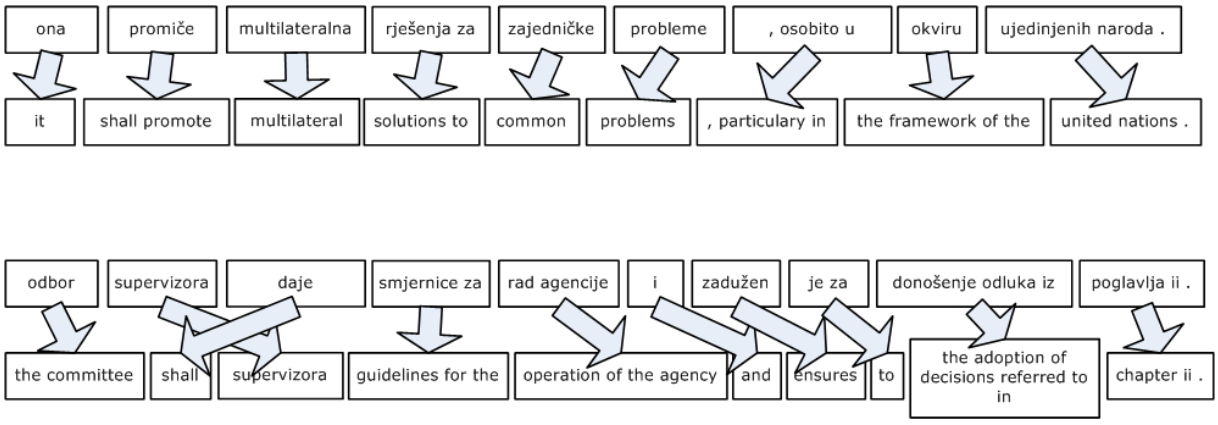

Figure 1: Illustration of the Croatian-English PBSMT.

are given at the end of section four. The major findings are summarized in the conclusion.

2 Related work

SMT systems started to be enriched with simple linguistic transformations already in 1992 [3]. Pseudo-lemmatization is one type of morphological analysis that is conducted in the pre-processing phase. It changes surface forms of words into new forms, either lemmas, roots, part-of-speech (POS) tags and morphemes, or the combination of these. According to [32], a morphological analysis should be done when there is no sizeable parallel corpus. Moreover, the authors show that lemmatization as one possible morphological analysis has no effect only when used on medium sized corpus. Lee [15] also shows that morphological analysis boosts translation quality even when there is a sizeable parallel corpus.

The research presented in this paper is similar to the research in Goldwater and McClosky [9]. They show that SMT systems can be improved by conducting morphological analysis of the source language training and test sets. Three of the suggested methods imply only pre-processing changes. The first suggested method is lemmatization which decreases the total number of distinct tokens. The authors experiment with lemmatizing only certain POS categories and with lemmatizing only scarce word forms. The latter modification shows promising results on the tuning set. Lemmatizing by truncating word forms, although beneficial when compared to the approach without lemmatization, proves to bring statistically significant deteriorations in scores compared to full lemmatization. The second suggested method introduces pseudo-tags which stand for grammatical categories. The third method is pseudo-lemmatization, which differs from pseudo-tagging in that the lemma and the tag are written as a unit. It proves helpful in cases where a phenomenon is expressed morphologically in both languages. The "simplification" of morphology is also present in Watanabe et al. [31], where stemming is done prior to truncating to four letters.

The problem of small parallel corpus can be tackled with phrasal lexicon, as given in Popović and Ney [23].

The related work involving south Slavic languages can be found in Popović et al. [21, 22], Popović and Ney [23], and Maučec and Kačić [17]. Popović et al. [21] exploit selected stemming in the Serbian-English translation direction. Popović et al. [22] conduct two types of morphological analysis, similarly to Goldwater and McClosky [9]. These entail lemmatization and inserting person pseudo-tags before verbs. Both studies lead to error reduction. In a subsequent study, Popović and Ney [23] show that article exclusion is beneficial in the English-Serbian set-up. They emphasize that enriching verbs with person pseudo-tags helps, but only if the corpus is not too small. Lemmatization as a pre-processing step is also applied in [17] for the Slovenian-English language pair and error reduction is achieved.

SMT involving Croatian is explored in [16] and [2]. The former entails the weather forecast domain and the latter the legal domain. Both studies involve only Croatian-English translation direction.

3 Experiments

3.1 Language pair

The languages selected for this research belong to two different categories from multiple aspects - while English is resource rich and morphologically poor, Croatian is resource poor and morphologically rich language. In this paragraph only major characteristics of Croatian are enlisted, for the reader to gain insight into the differences that might hinder the translation process. Croatian belongs to a Slavic group of languages, which are characterized by a relatively free word order. English, on the other hand, is a subject-verb-object lan-

guage. Croatian inflectional morphology is rich for all open word classes. Nouns, pronouns, adjectives, and some numbers are fully inflected with seven grammatical cases, two grammatical numbers and three grammatical genders. There are seven cases with a simplification developed over time according to which dative and locative of singular, and dative, locative and instrumental of plural, and nominative and vocative of plural are associated to the same word forms [8]. In English, pronouns are inflected only with three grammatical cases, while nouns change form only in the possessive case. Gender, on the other hand, is not an inflectional category. The grammatical number is displayed in nouns, pronouns and articles. In Croatian, verbs are conjugated to communicate gender, person, number, tense, aspect, mood and voice, while English verbs are not heavily inflected. In general, the only inflected forms of English verbs are third person singular present tense forms, past tense forms, past participle (which may be the same as the past tense) forms, and -ing forms that serve as present participles and gerunds. Furthermore, some Croatian nouns have two different plural forms depending on the number preceding them [30]. With verbs, person is often expressed though suffix and pronoun is often dropped. The negation of the three most important verbs (to be, to have and to want) is formed by affixing the particle *ne* [30]. The lack of articles in Croatian is another major difference between the two languages concerned.

Since the contrastive analysis of the languages reveals many differences, a morphological analysis is made prior to the second part of the research.

3.2 Language resources

A small subset of the English-Croatian part of the Acquis Communautaire is checked for spelling mistakes, manually segmented, and manually segment-aligned. We refer to the segment, rather than the sentence, because representative sentences of Acquis Communautaire are in most cases either extremely short or extremely long. Manual work is done using the tool *CorAl* described in [25]. In order to reduce the number of distinct tokens, all the digits are replaced with a token @. The set is tokenized, lowercased, and cleaned from redundant spaces or empty lines. The segment-aligned parallel corpus, which entails 300 documents from 1957 to 2007, forms the translation training set (Table 1). All the rows in Table 1 and subsequent tables which show the number of tokens, unique tokens, and *hapax legomena*, do not include punctuation, and the statistics are calculated after replacing the digits with @. There is an obvious difference in the numbers of tokens and distinct tokens between Croatian and English part of the parallel corpus, the reasons for which lie in the language divergences discussed in the previous section. Rich morphology decreases the total number of tokens but increases the number of distinct tokens. Finally, all the segments which have over 100 tokens or which do not comply with the 9:1 ratio imposed by GIZA++ tool are filtered out. Frequency analysis reveals that the most frequent Croatian word is the preposition *u*, while the most frequent English word is the article *the*. In general, the majority POS category in the top ten frequency list in both languages are prepositions.

Table 1: Filtered training set.

	CRO	**EN**
# of segments	35467	35467
avg seg length	14	16
# of tokens	569839	661160
# of unique tokens	24534	10990
# of *hapax legomena*	9364	3493

The tuning set (Table 2) and the two tests sets (Table 3, 4) are based on 11 documents created from 2007 onwards and prepared in the same way. Each contains the total of 638 sentences. All the sets except for the translation training set and language training set are disjunct.

Table 2: Tuning set statistics.

Tuning set		
	CRO	EN
# of sentences	638	638
avg sentence length	20	25
# of tokens	14631	17706
# of unique tokens	3329	2123
# of *hapax legomena*	1848	975

Table 3: Test set 1 statistics.

Test set 1		
	CRO	EN
# of sentences	638	638
avg sentence length	21	25
# of tokens	14715	17908
# of unique tokens	3268	2163
# of *hapax legomena*	1789	903]

The English language model is trained on the English version of Acquis Communautaire [27]. The total of 23.545 documents is collected (Table 5). Since the documents contain closing phrases in all of the official EU languages, these phrases have been filtered out in

Table 4: Test set 2 statistics.

Test set 2		
	CRO	EN
# of sentences	638	638
avg sentence length	21	26
# of tokens	15106	18653
# of unique tokens	3227	2025
hapax legomena	1713	884

all but the English language by running a script after manually inspecting the documents.

Table 5: Language model training set statistics.

Language model training set	
# of sentences	2389943
avg sentence length	24
# of tokens	65016878
# of unique tokens	326438
# of *hapax legomena*	139896

3.3 Croatian-English SMT systems

All the models built within this research have fourteen standard features - 3-gram language model score [28], phrase translation table scores in both directions, lexical weighing scores in both directions (the product of lexical probabilities of words aligned to corresponding foreign language words within phrases), phrase penalty, word penalty, and seven reordering features. Besides the distance model which punishes movements as they get larger, the remaining reordering features are the lexical model scores in both directions, whereas only three reordering types are taken into consideration - monotone, swap with previous phrase, and discontinuous. An open source tool *Moses* is used for training the model components other than the language model component [12].

An interpolated trigram language model is trained with Chen-Goodman modification of Kneser-Ney smoothing. Since *in vivo* evaluation of a language model would be too expensive, an intrinsic evaluation is performed, which uses perplexity as a metric [10]. There are about 3-4 percent of unknown words in each of the sets before the replacement of digits with the token @, and about 1 percent after the replacement.

One SMT system (LegTran A) is built with the above described resources. Prior to training the other system (LegTran B), the Croatian part of the parallel corpus is partially pseudo-lemmatized by combining and modifying the first and the third method defined in [9]. All the distinct tokens that have a frequency of two are taken into consideration for pseudo-lemmatization. *Hapax legomena* are not taken into account because they are three times more numerous and they often result from the tokenization error. Out of this preselected set of tokens, only those tokens the root of which appears at least twice in the set are selected for pseudo-lemmatization. Lemmatization is done with the help of the Croatian Lemmatization Server [29] and with the following limitations:

- number category with nouns - *djeci* (children) is pseudo-lemmatized as *djeca* and not as *dijete* (child), i.e. plural nouns are reduced to plural nominative, and not to singular nominative as in full lemmatization
- degree with adjectives - *široj* (wider) is pseudo-lemmatized as *širi* and not as *širok* (wide), i.e. adjectives are reduced to nominative of the corresponding degree, and not to positive as in full lemmatization
- tense tag is added to the present participle - *predložio* (proposed) is pseudo-lemmatized as *predložitiS* and not as *predložiti* (propose)
- third person singular verbs in present tense - *predlaže* (proposes) is not pseudo-lemmatized
- homographs - *sam* is not pseudo-lemmatized because it can stand for the weak form of the first person singular present tense helping verb 'to be' or for the adjective 'alone', i.e. words that share the same written form but differ in meanings are not pseudo-lemmatized

The number of distinct tokens is reduced by 2 percent, while the number of *hapax legomena* is reduced by the total of 51, or 0.01 percent. The highest number of changes is done on adjectives (Table 6).

Table 6: The number of pseudolemmatized forms per POS.

	N	**V**	**Adj**	**P**
training set	163	146	1148	6
tuning set	6	1	31	0
test set 1	1	1	21	0
test set 2	5	1	35	0

4 Results

Automatic evaluation is performed using the most widespread metrics in SMT - TER, BLEU and Meteor. Significance testing is performed using approximate randomization and the sign test, which, unlike

bootstrapping [24], are not susceptible to type I mistakes [6].

4.1 Without optimization

In the first part of the evaluation, optimization of system parameters is not performed, i.e. system parameters are set to default values - reordering models get the weight of 0.3, language model 0.5, translation models 0.2, and word penalty gets the weight of -1. Motivation for such an experiment is found in Ozdowska and Way [20]. Although optimization would definitely lead to better absolute overall results [7], there is no reason to believe that it would give radically different relative results. We wanted to confirm that intuition.

BLEU difference between the system without pseudo-lemmatization and the one with pseudo-lemmatization is not statistically significant according to approximate randomization. Furthermore, since there is a huge variance in each test set result, the test sets are joined, which results in lower variance and therefore confirms the necessity of having a bigger test set.

Since automatic evaluation leads to no significant conclusions, a quick manual analysis of the pseudo-lemmatization effects is done. The total of 10 sentences, which are affected by pseudo-lemmatization, are extracted and errors are analyzed. A short look at the differences in translations obtained by LegTranA and LegTranB is given in Table 7. SRC_A stands for the source sentence provided to the LegTranA, and SRC_B stands for the source sentence provided to the LegTranB. The first sentence illustrates that the word *kombinaciju* is correctly translated only after pseudo-lemmatization. The second sentence illustrates the same point with the verb *provelo*. The last sentence illustrates the surrounding word translation changes caused by pseudo-lemmatization.

Altogether, in all the analyzed sentences one verb pseudo-lemma actually deteriorates the translation, while one verb pseudo-lemma makes the translation possible. With nouns, cases with the same the translation and cases with the facilitated translation are detected. With adjective pseudo-lemmas, in one of the selected sentences the translation deteriorates, while in other three cases the translation stays the same or the adjective stays untranslated but the surrounding word translations improve.

This short analysis suggests that pseudo-lemmatization has some positive effects and that these effects would possibly be reflected in automatic scores if both translations, reference and machine, were stemmed. Lavie et al. [14] show that automatic and human scores correlate better when more weight is given to recall than to precision. This is due to the fact that translation may be recovered but not in the exact form, which metrics like BLEU do not appreciate as they should. Although pseudo-lemmatization recovers certain translations, not all of them add up to the final BLEU score. More precisely, only those that are recovered in the exact form influence the automatic evaluation result. Human evaluation, on the other hand, values all of them as far as the criterion of adequacy is concerned.

4.2 With optimization

After running MERT optimization on the tuning set for both systems, BLEU results of each system individually improve between 0.6 and 0.8 BLEU points on both test sets. The differences in scores prove to be statistically significant according to approximate randomization. Since MERT is known for its instability, we conduct optimization at least 3 times in order to compare the optimized systems properly. The difference in BLEU, Meteor and TER results between LegTran B and LegTran A on test set 1 and joint test set is statistically significant, while the one on test set 2 is significant only for TER (Table 8).

Table 8: BLEU, Meteor and TER results on test sets translated by MERT-optimized systems marked by A and B, whereas A stands for LegTran A and B stands for LegTran B.

		BLEU↑	Meteor↑	TER↓
test set 1	A	**38.8**	**35.8**	**46.1**
	B	**39.2**	**36.0**	**45.7**
test set 2	A	38.2	35.9	**46.6**
	B	38.3	35.9	**46.0**
test sets	A	**38.5**	**35.8**	**46.4**
	B	**38.7**	**36.0**	**45.9**

The best translations of a small subset of source sentences, i.e. the closest to their respective reference translations (marked as REF) in terms of BLEU, are provided for illustration purposes in Table 9.

4.3 Manual evaluation

Manual evaluation is conducted according to the criteria of adequacy and fluency on a typical 1-5 scale. The evaluation is performed by two professional translators with a degree in English. Altogether 400 randomly chosen sentences translated by LegTran A and 400 randomly chosen sentences translated by LegTran B are evaluated. The judges are oblivious to the origin of translations and each of them judges sentences from both systems interchangeably in a completely random order. The results are given in Table 10.

There is the total of 130 overlapping source sentences judged for both systems. The sign test checks how likely a sample of better and worse BLEU scores would have been generated by two systems of equal

Table 7: LegTranA and LegTranB translation differences.

SRC_A	podnositelj zahtjeva odabire jedan modul ili kombinaciju modula koji su navedeni u sljedećoj tablici .
SRC_B	podnositelj zahtjeva odabire jedan modul ili kombinacija modula koji su navedeni u sljedećoj tablici .
LegTranA	the applicant shall select one modul or kombinaciju of module referred to in the following table .
LegTranB	the applicant shall select one modul or a combination of module referred to in the following table .
SRC_A	ono podnositelju (podnositeljima) zahtjeva izdaje ispitno izvješće i po potrebi izvješće o inspekcijskom pregledu i / ili testiranju , ako ga je provelo .
SRC_B	ono podnositelju (podnositeljima) zahtjeva izdaje ispitno izvješće i po potrebi izvješće o inspekcijskom pregledu i / ili testiranju , ako ga je provesti1S .
LegTranA	it by the (podnositeljima) applications shall be issued by a test report and , where appropriate , a report on the inspekcijskom pregledu and / or a test , if by .
LegTranB	the council by (podnositeljima) applications shall be issued by a test report and , where appropriate , a report on the inspekcijskom pregledu and / or a test , if it is applied .
SRC_A	svaki nacrt zakonodavnog akta mora sadržavati detaljnu izjavu kojom se omogućuje procjena poštovanja načela supsidijarnosti i proporcionalnosti .
SRC_B	svaki nacrt zakonodavnog akta mora sadržavati detaljan izjavu kojom se omogućuje procjena poštovanja načela supsidijarnosti i proporcionalnosti .
LegTranA	each of the zakonodavnog act must contain the detailed declaration which allow the assessment of respect the principle of subsidiarity and proportionality .
LegTranB	any draft zakonodavnog act must include a detailed statement of assurance as to allow the assessment of respect the principle of subsidiarity and proportionality .

Table 9: The best translations for a selected set of sentences.

SRC	vijeće , djelujući jednoglasno na prijedlog komisije , donosi odluku kojom se određuje sastav odbora .
BEST	the council , acting unanimously on a proposal from the commission , shall adopt the decision determining the composition of the committee .
REF	the council , acting unanimously on a proposal from the commission , shall adopt a decision determining the committee 's composition .
SRC	vijeće djeluje u skladu s člankom @@@.e.
BEST	the council shall act in accordance with article @@@.e.
REF	the council shall act in accordance with article @@@ e .
SRC	ovom se izjavom ne dovode u pitanje odredbe ugovora kojima se uniji dodjeljuje nadležnost , uključujući i u području socijalnih pitanja .
BEST	this declaration shall be without prejudice to the provisions of the union jurisdiction , including in the field of social issues .
REF	this declaration is without prejudice to the provisions of the treaties conferring competence on the union , including in social matters .
SRC	taj je postupak utvrđen člankom @@@ .
BEST	this procedure laid down in article @@@ .
REF	this procedure is defined in article @@@ .

Table 10: Manual evaluation results of optimized systems according to the criteria of fluency and adequacy.

	Sys	Fluency	Adequacy	Total
Grades	A	3.55	3.28	6.83
	B	3.40	3.51	6.91

performance. According to the sign test LegTran B is significantly better with respect to the adequacy criterion. No conclusion can be reached for the second criterion.

A strong positive correlation is determined between the two criteria according to the Pearson correlation coefficient. Furthermore, a moderate negative correlation between the number of words and fluency and a weak negative correlation between the number of words and adequacy grades is determined in LegTranB.

5 Conclusion

Partial pseudo-lemmatization on the Croatian part of the parallel corpus brings slight improvements. However, if only automatic evaluation were performed on the non-optimized systems, these improvements would remain unknown. On the MERT optimized systems, the score improvements achieved with pseudo-lemmatization reflect statistically significant differences on the first and joint test sets according to all three metrics. Due to the instability of the optimization procedures undertaken, conclusions are made with respect to the important parameters. However, prior to reaching any conclusions on pseudo-lemmatization based on automatic evaluation results, effects of different optimization techniques need to be thoroughly examined.

Manual evaluation gives insight into the nature of the differences achieved and reveals that pseudo-lemmatization boosts adequacy at the expense of fluency. The effects of pseudo-lemmatization still need to be confirmed in a bigger scenario incorporating many more pseudo-lemmas. Another interesting line of research would be to examine pseudo-lemmatization effects as the training set size grows. Finally, since pseudo-tag for person shows to be important for Czech which is a pronoun drop language, in our future work we plan to incorporate that set of pseudo-tags.

6 Acknowledgment

This research has been supported under the Grant No. 13.13.1.3.03 of the University of Rijeka.

References

[1] Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65-72, 2005.

[2] Brkić, M; Bašić Mikulić, B; Matetić, M. Can we beat Google Translate? In *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces (ITI)*, pages 381-386, 2012.

[3] Brown, P.F; Della Pietra, S.A; Della Pietra, V.J; Lafferty, J.D; Mercer, R.L. Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 83-100, 1992.

[4] Brown, P.F; Pietra, V.J.D; Pietra, S.A.D; Mercer, R.L. A statistical approach to machine translation. *Computational linguistics*, 16(2):79-85, 1990.

[5] Brown, P.F; Pietra, V.J.D; Pietra, S.A.D; Mercer, R.L. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263-311, 1993.

[6] Clark, J.H; Dyer, C; Lavie, A; Smith, N.A. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT)*, Vol. II, pages 176-181, 2011.

[7] Foster, G; Kuhn, R; and others. Stabilizing minimum error rate training. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 24-249, 2009.

[8] Gis, I. Dativ i lokativ u suvremenim gramatikama. *Hrvatistika*, 5(5):47-55, 2011.

[9] Goldwater, S; McClosky, D. Improving statistical MT through morphological analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT EMNLP)*, pages 676-683, 2005.

[10] Jurafsky, D; Martin, J.H. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2009.

[11] Koehn, P; Och, F.J; Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL HLT)*, Vol. I, pages 48-54, 2003.

[12] 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; Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177-180, 2007.

[13] Koehn, P. *Statistical machine translation*. Cambridge University Press, Cambridge, UK, 2010.

[14] Lavie, A; Sagae, K; Jayaraman, S. The significance of recall in automatic metrics for MT evaluation. *Machine Translation: From Real Users to Research*, pages 134-143, 2009.

[15] Lee, Y.S. *Morphological analysis for statistical machine translation*. Defense Technical Information Center. 2004.

[16] Ljubešić, N; Bago, P; Boras, D. Statistical machine translation of Croatian weather forecasts: How much data do we need? *Journal of Computing and Information Technology*, 18(4), 2011.

[17] Maučec, M.S; Kačić, Z. Statistical machine translation from Slovenian to English. *Journal of Computing and Information Technology*, 15(1):47-59, 2007.

[18] Och, F.J; Ney, H. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19-51, 2003.

[19] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311-318, 2002.

[20] Ozdowska, S; Way, A. Optimal bilingual data for French-English PB-SMT, 2009.

[21] Popović, M; Jovičić, S.T; šarić, Z.M. Statistical machine translation of Serbian-English. In *Proceedings of the Ninth Conference Speech and Computer*, 2004.

[22] Popović, M; Vilar, D; Ney, H; Jovičić, S; šarić, Z. Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 41-48, 2005a.

[23] Popović, M; Vilar, D; Ney, H; Jovičić, S; šarić, Z. Augmenting a small parallel text with morpho-syntactic language resources for Serbian-English statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 2005b.

[24] Riezler, S. and Maxwell, J.T. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57-64, 2005.

[25] Seljan, S; Tadić, M; Agić, ž; šnajder, J; Dalbelo Bašić, B; Osmann, V. Corpus Aligner (CorAl) evaluation on English - Croatian parallel corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 3481-3484, 2010.

[26] Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*, pages 223-231, 2006.

[27] Steinberger, R; Pouliquen, B; Widiger, A; Ignat, C; Erjavec, T; Tufis, D. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006.

[28] Stolcke, A. SRILM-an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, 2002.

[29] Tadić, M. Croatian Lemmatization Server. In *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan languages*, 2006.

[30] Težak, S; Babić, S. *Gramatika hrvatskog jezika: priručnik za osnovno jezično obrazovanje*. 9. popravljeno izdanje. školska knjiga, Zagreb, 1996.

[31] Watanabe, T; Suzuki, J; Tsukada, H; Isozaki, H. NTT statistical machine translation for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 92-102, 2006.

[32] Zhang, R; Sumita, E. Boosting statistical machine translation by lemmatization and linear interpolation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 181-184, 2007.