Language related issues for machine translation between closely related South Slavic languages

Maja Popović[1] **Mihael Arčan**[2] **Filip Klubička**[3]

[1] Humboldt University of Berlin, Germany
maja.popovic@hu-berlin.de

[2] Insight Centre for Data Analytics, National University of Ireland, Galway
mihael.arcan@insight-centre.org

[3] Department of Information and Communication Sciences, University of Zagreb, Croatia
fklubick@ffzg.hr

Abstract

Machine translation between closely related languages is less challenging and exhibits a smaller number of translation errors than translation between distant languages, but there are still obstacles which should be addressed in order to improve such systems. This work explores the obstacles for machine translation systems between closely related South Slavic languages, namely Croatian, Serbian and Slovenian. Statistical systems for all language pairs and translation directions are trained using parallel texts from different domains, however mainly on spoken language i.e. subtitles. For translation between Serbian and Croatian, a rule-based system is also explored. It is shown that for all language pairs and for both translation systems, the main obstacles are the differences between syntactic properties.

1 Introduction

Machine translation (MT) between (closely) related languages is a specific field in the domain of MT which has attracted the attention of several research teams. Nevertheless, it has not attracted as much attention as MT between distant languages. This is, on the one side, due to the fact that speakers of these languages often easily understand each other without switching to the foreign language. Furthermore, many documents are distributed in their original language, even in the neighbouring countries. Another fact is that MT between related languages is less problematic than between distant languages (Kolovratník et al., 2009).

Still, there is a need for translation even between very closely related language pairs such as Serbian and Croatian, for example, for the sake of producing standard official documents which exixst in one language but not the other. Another application of such systems is the two-stage (also called "pivot") MT (Babych et al., 2007): for example, if an adequate English-Croatian system is available whereas an English-Serbian system is not, or is of poor quality, English source sentences can first be translated into Croatian, and then the obtained output is further translated into Serbian by a Croatian-Serbian MT system. A similar application can also include enriching parallel training corpora by producing "synthetic" data in the less resourced related language (Bertoldi and Federico, 2009).

This work examines MT systems between three closely related South Slavic languages, namely Croatian, Serbian and Slovenian. Therefore we used the *Asistent*[1] phrase-based translation system (Arčan et al., 2016), which was developed to translate text between English and the morphological complex south Slavic languages: Slovene, Serbian and Croatian. Additionally, an RBMT system[2] (Klubička et al., 2016) is analysed for translation between Croatian and Serbian in both directions in order to explore advantages and disadvantages of both approaches for very close language pairs.

[1] http://server1.nlp.insight-centre.org/asistent/

[2] http://translator.abumatran.eu

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 43–52, Osaka, Japan, December 12 2016.

Research questions

Taking into account the language differences among Croatian, Serbian and Slovenian, our main questions are:

- What are the main obstacles for machine translation between these languages?
- Considering the closeness between Serbian and Croatian, which approach exhibits fewer errors, SMT or RBMT? What are the most important differences between the two approaches?

1.1 Related work

Although all South Slavic languages are still rather under-resourced and under-investigated, in the last decade several MT systems have been built between these languages and English. Nevertheless, the translation between them has been investigated to a much lesser extent.

A rule-based translation system between Slovenian and Serbian has been described in Vičič (2008) and automatic scores (BLEU, METEOR and edit-distance) as well as adequacy and fluency are reported. Another work on RBMT between Serbian, Croatian and Slovenian is presented in Peradin et al. (2014). The SUMAT project[3] included a statistical approach for Serbian and Slovenian subtitles (Etchegoyhen et al., 2014). Nevertheless, a deeper analysis of translation errors or problems has not been performed in any of these articles.

Evaluation of several scenarios with different models and data-sets involving Croatian are explored in Toral et al. (2016) in the framework of the Abu-MaTran project[4], but only for translation from and to English. Three MT systems between Croatian and Serbian for the news domain are described in Popović and Ljubešić (2014), one very basic rule-based system and two SMT systems trained on small and on large parallel texts. Their performance was not examined in detail however, as they are only used as a bridge for translation from and into English.

Analysis of problems for MT between closely related languages together with a comparison between an RBMT and an SMT system is presented in Kubon and Vičič (2014) for the Czech-Slovak language pair. Similar analysis for South Slavic languages has been performed in Popović and Arčan (2015), though not for translation between these languages but from and into English and German.

To the best of our knowledge, no systematic investigation of actual difficulties for MT systems translating between South Slavic languages has been carried out yet.

2 Language properties – similarities and differences

2.1 Common properties

All three languages, Croatian, Serbian and Slovenian, belong to the South-Western Slavic branch. As Slavic languages, they have a very rich inflectional morphology for all word classes. There are six distinct cases affecting not only common nouns, but also proper nouns as well as pronouns, adjectives and some numbers. Some nouns and adjectives have two distinct plural forms depending on the number (less than five or not). There are also three genders for the nouns, pronouns, adjectives and some numbers leading to differences between the cases and also between the verb participles for past tense and passive voice. When it comes to verbs, person and many tenses are expressed by the suffix, and, similarly to Spanish and Italian, the subject pronoun (e.g. *I, we, it*) is often omitted. In addition, negation of three quite important verbs, *biti* (all languages) (*to be*), *imati* (Croatian, Serbian) / *imeti* (Slovenian) (*to have*) and *ht(j)eti* (Croatian, Serbian) / *hoteti* (Slovenian) (*to want*), is formed by adding the negative particle to the verb as a prefix. There are also two verb aspects, and so many verbs have perfective and imperfective form(s) depending on the duration of the described action. The different forms are lexicalized, and are often either different but very similar (e.g. *skakati-skočiti*), or are distinguished only by prefix (e.g. *gledati-pogledati*). It should be noted that this phenomenon is less prominent in Slovenian.

As for syntax, all three languages have quite a free word order, and neither language uses articles, either definite or indefinite. In addition to this, multiple negation is always used.

[3] http://www.sumat-project.eu

[4] http://www.abumatran.eu/

It should be also noted that while the Latin alphabet is common for all three languages, Serbian also uses the Cyrillic script. However, this poses no problem regarding MT because a Cyrillic Serbian text can be easily transliterated into Latin, as there is one-to-one correspondence between the characters.

2.2 Differences between Croatian and Serbian

Croatian and Serbian exhibit a large overlap in vocabulary and a strong morpho-syntactic similarity so that the speakers can understand each other without difficulties. Nevertheless, there is a number of small but notable and also frequently differences occurring differences between them.

The largest differences between the two languages are in vocabulary: some words are completely different, some however differ only by one or two letters. In addition, Serbian language usually phonetically transcribes foreign names and words although both transcription and transliteration are allowed, whereas the Croatian standard only transliterates.

Apart from lexical differences, there are also structural differences mainly concerning verbs: modal verb constructions, future tense, conditional, as well as constructions involving the verb *trebati* (to need, should). When it means *should*, in Croatian it takes the tense according to the subject and it is transitive as in English (*trebam raditi* equals *I should work*). In Serbian however, it is impersonal followed by the conjunction da and the present of the main verb (*treba da radim* equals *I should work*). When it means *to need*, the Croatian structure is the same (*trebam posao* equals *I need a job*, *Petar treba knjige* equals *Petar needs books*), whereas in Serbian, the verb is conjugated according to the needed object, and the subject which needs something is an indirect grammatical object in dative case (*meni treba posao = I need a job*, *Petru trebaju knjige = Petar needs books*). The Serbian structure is also possible in Croatian, although the other one is preferred. Impersonal constructions (*treba uraditi = it should be done*) are same in both languages, namely the verb *trebati* in third person singular is followed by infinitive of the main verb.

Regarding other modal verbs, the infinitive is prescribed in Croatian (*moram raditi = I have to work*), whereas the construction with conjunction *da* (en. *that/to*) and present tense is preferred in Serbian (*moram da radim*). The mentioned difference partly extends to the future tense which is formed in a similar manner to English, i.e. using present of the verb *ht(j)eti* as the auxiliary verb. The infinitive is formally required in both variants, however, when *da*+present is used instead, it can additionally express the subject's will or intention to perform the action. This form is frequent in Serbian (*ja ću da radim = I will work*), whereas in Croatian only the infinitive form is used (*ja ću raditi*). Another difference regarding future tense exists when the auxiliary and main verb are reversed: in Croatian, the final *i* of the infinitive is removed (*radit ću*), whereas in Serbian the main and the auxiliary verb merge into a single word (*radiću*).

2.3 Differences from Slovenian

Even though Slovenian is very closely related to Croatian and Serbian, and the languages share a large degree of mutual intelligibility, a number of Croatian/Serbian speakers may have difficulties with Slovenian and the other way round.

The nature of the lexical differences is similar to the one between Croatian and Serbian, namely a number of words is completely different and a number only differs by one or two letters. However, the amount of different words is much larger. In addition to that, the set of overlapping words includes a number of false friends (e.g. *brati* means *to pluck* in Croatian and Serbian but *to read* in Slovenian).

The amount of grammatical differences is also larger and includes local word order, verb mood and/or tense formation, question structure, dual in Slovenian, usage of some cases, structural properties for certain conjunctions as well as some other structural differences. Local word order differences include, for example, the order of auxiliary and main verbs: Slovenian allows the auxiliary verb to be at the beginning of the clause, whereas Croatian and Serbian do not (*sem videl/videl sem = video sam = I've seen*). Also, the place of reflexive pronoun is different (*se vidi = vidi se = it can be seen*, *se mi zdi = čini mi se = it seems to me*).

Constructions involving the Croatian/Serbian verb *trebati* differ significantly: in Slovenian, the meaning *should* is expressed by the adverb *treba* (*bi bilo treba = trebalo bi = it should*). For the meaning *to*

need, the verb *potrebovati* is used in the same form as the verb *trebati* in Croatian, i.e. it requires the needed object in accusative case (*potrebujem knjigo = trebam knjigu = I need a book*).

The main difference regarding tense formation is the future tense. In Slovenian, it is formed using the auxiliary verb *biti* and the past participle of the main verb – in Croatian and Serbian, another auxiliary verb is used, *ht(j)eti* with the infinitive or *da* + present tense of the main verb (*jaz bom videl = ja ću da vidim = ja ću vid(j)eti = I will see*). Another important difference is Slovenian conditional formed using the adverb *lahko* and present tense of the main verb: in Croatian and Serbian it is formed by the modal verb *moći* (*can* and infinitive or *da* + present tense (*lahko vidim = mogao bih da vidim = mogao bih videti = I could see*).

Some conjunctions and/or require completely different structuring. For example, Slovenian *tudi* (en. *also*, *too*) has a direct equivalent in Croatian and Serbian (*takodje(r)*), but it is often translated by *i*. For negation form *neither* in Slovenian the construction *tudi ne* is used, whereas in Croatian and Serbian a negation conjunction *ni* is used. Slovenian conjunction *pa* also has different usage and structural requirements, and it can also be considered as a false friend.

Another important difference is the Slovenian dual grammatical number which refers to two entities (apart from singular for one and plural for more than two). It requires additional sets for noun, adjective and verb inflexion rules not existing either in Croatian or in Serbian.

3 Experimental set-up

3.1 Machine translation systems

The statistical phrase-based systems (Koehn, 2004) were trained using the Moses toolkit (Koehn et al., 2007) with MERT tuning. The word alignments were built with GIZA++ (Och and Ney, 2003) and a 5-gram language model was built with kenLM (Heafield, 2011). The parallel texts used to train the SMT systems were mostly obtained from the OPUS[5] web site (Tiedemann, 2009), which contains various corpora of different sizes and domains. Although corpora in distinct domains, e.g., legal, medical, financial, IT, exist for many language pairs including some of the South Slavic languages and English, parallel data between South Slavic languages pairs consist mostly of the OpenSubtitles[6] corpus and a little portion of the technical domain. For Serbian-Croatian language pair, the SEtimes corpus from the news domain (Tyers and Alperen, 2010) is also available. In total, about 15 million of sentence/segment pairs containing about 100 million of running words was used for training (Table 1). For tuning, 2000 sentence pairs were used for each language pair.

The Croatian-Serbian RBMT system is a bidirectional rule-based system which is based on the open-source Apertium platform (Forcada et al., 2011) and has been built collaboratively between several institutions as part of the aforementioned Abu-MaTran project. The process involved several workshops that employed the work of experts and non-experts to gather the necessary data to build a bilingual dictionary and to verify correct transfer rules automatically inferred using a tool developed by Sánchez-Cartagena et al. (2015). Work on the translator has continued since, and at the time of writing this paper the bilingual dictionary has quite a high coverage, containing a total of 88521 bilingual lemma entries, while the number of defined transfer rules in the Serbian-Croatian direction is 99, and 86 in the Croatian-Serbian direction. At the time of publication, the system was automatically evaluated on 351 Serbian sentences gathered from newspaper texts that were manually translated into Croatian, and when compared to *Google Translate*, the only other available system at the time, the RBMT system yielded higher scores. For more details on the construction and evaluation of the system, refer to Klubička et al. (2016).

3.2 Test sets

The in-domain data set used for evaluating SMT performance consists of about 2000 sentences for each language pair isolated from the training data set. Therefore, the test data consist mostly out of the OpenSubtitles corpus, since this corpus builds the largest part (95%) of the data used to train the translation models.

[5] http://opus.lingfil.uu.se/

[6] http://www.opensubtitles.org

Corpus Name	Slovene-Croatian	Slovene-Serbian	Croatian-Serbian
Gnome	4K	600K	300K
KDE	85K	49k	33.2k
OpenSubtitles	6.1M	13.3M	22.3M
SETimes	/	/	200K
Ubuntu	557	86K	51K
Training Data	Sl-Hr	Sl-Sr	Hr-Sr
L1 words:	39M	90M	137M
L2 words:	40M	94M	139M
unique L1 w.:	468K	775K	1.22M
unique L2 w.:	579K	966K	1.24M
Par. sentences:	5.5M	12.6M	19.4M

Table 1: Statistics on parallel corpora used to build the translation models (explanation: Slovene-Croatian → L1=Slovene, L2=Croatian).

Such data sets are usual for evaluation and comparing SMT systems, however, they are not optimal for comparing an SMT and an RBMT system since they originate from the same text type as the SMT training corpus – the results would probably be biased. Therefore, additional test sets were created for this comparison:

- 1000 Croatian source sentences were extracted from the hrenWaC and DGT part of the OPUS data and translated by both systems into Serbian; about 300 segments from each of the translation outputs were post-edited by native speakers.
- 3000 Serbian source sentences were extracted from a corpus containing language course material and translated by both systems into Croatian; about 450 segments from each of the translation outputs were post-edited by native speakers.

In addition, a subset of the Slovenian-to-Serbian SMT translation output containing about 350 sentences was post-edited as well.
The test sets were post-edited for two reasons:

1. post-edited data are generally more convenient for analysis and identifying prominent errors and issues;
2. the OpenSubtitles contain translations from English the as original source so that the obtained translations are often too different and do not fully reflect the language closeness.

Although it was not the motivation for post-editing, it should be noted that there were no available reference translations for Croatian-Serbian additional test sets.

3.3 Evaluation

For all test sets and MT systems, BLEU scores (Papineni et al., 2002) and character n-gram F-scores CHRF3 (Popović, 2015) are reported. BLEU is a well-known and widely used metric, and CHRF3 is shown to correlate very well with human judgments for morphologically rich languages (Stanojević et al., 2015). Besides, it seems convenient for closely related languages since a large portion of differences is on the character level.

In order to better understand the overall evaluation scores and differences between the MT systems, five error classes, produced by the automatic error analysis tool Hjerson (Popović, 2011), are reported.

Finally, in order to determine most prominent language related issues for the MT systems, a manual inspection of the errors and their causes is carried out, predominantly on the post-edited data.

	BLEU	CHRF3
Serbian→Croatian	70.1 (64.9)	80.9 (78.6)
Croatian→Serbian	67.4 (59.9)	78.0 (73.8)
Serbian→Slovenian	29.2 (14.1)	47.4 (37.2)
Slovenian→Serbian	23.5 (12.3)	43.2 (34.3)
Croatian→Slovenian	38.6 (16.1)	55.0 (39.5)
Slovenian→Croatian	34.6 (13.5)	51.0 (37.4)

Table 2: Automatic translation scores BLEU and CHRF3 for the SMT system (together with the *Google Translate* system in parentheses) on the in-domain test set.

	inflection	order	omission	addition	lexical	ΣERR
Serbian→Croatian	1.8	1.3	3.6	4.9	12.0	23.5
Croatian→Serbian	2.0	1.4	5.0	3.9	14.9	27.7
Serbian→Slovenian	3.7	4.5	10.1	12.1	27.6	58.0
Slovenian→Serbian	3.4	3.9	14.6	9.1	30.1	62.0
Croatian→Slovenian	3.1	4.2	8.8	11.4	24.2	51.7
Slovenian→Croatian	3.1	3.8	12.4	8.0	28.2	55.6

Table 3: Translation error classes of the SMT system identified by the Hjerson tool on the in-domain test set.

4 Evaluation results on standard in-domain test sets

Table 2 presents the automatic scores for standard test sets for all SMT systems together with the scores for translations[7] by the publicly available Google translate[8] system.

The obtained scores are rather high for translation between Serbian and Croatian and lower for translations involving Slovenian. Nevertheless, considering the language closeness, the scores are not particularly high – the most probable reason are the "unnecessary" differences introduced by human translation from a third language, namely English. It can be noted that translation into Serbian is worse than into the other two languages and that translation into Slovenian is better than into the other two languages.

Table 3 gives details on the translation error classes. For the Serbian-Croatian language pair most errors are lexical, whereas there is a rather low number of inflectional and ordering errors. This can be expected considering that the main differences between the languages are on the lexical level, as described in Section 2.2. As for translating from and into Slovenian, lexical errors are also predominant and much more frequent. Furthermore, the amount of ordering and inflectional errors is not negligible. These results are consistent with the language differences described in Section 2.3, however, they are not giving precise information of which phenomena are causing which errors. For this purpose, a shallow manual inspection of errors is carried out. It has been noted that the structural differences often result in different error types. Nevertheless it was not easy to isolate specific phenomena due to the described suboptimal test sets. Therefore the manual inspection of errors has been carried out thoroughly on the post-edited data and the results are reported in the next section.

5 Evaluation results on post-edited test sets

The first evaluation step of post-edited test sets was performed to calculate automatic evaluation metrics and class error rates. After that, a detailed manual inspection of language related phenomena leading to particular errors is carried out. Finally, the most problematic phenomena were isolated from test sets and evaluated separately.

[7] generated in September 2016

[8] https://translate.google.com/

		BLEU		CHRF3	
		SMT	RBMT	SMT	RBMT
Serbian→Croatian	overall	91.0	89.6	95.4	95.1
	trebati	52.4	54.8	77.5	78.5
Croatian→Serbian	overall	86.2	82.9	93.4	92.2
	trebati	58.8	62.5	83.4	84.4

		SMT		RBMT	
Serbian→Croatian		overall	*trebati*	overall	*trebati*
	ΣERR	4.3	29.6	4.8	27.6
	inflection	1.2	13.7	1.3	13.0
	order	0.3	0.2	0.4	0.0
	omission	0.4	0.0	0.0	0.0
	addition	0.9	5.5	0.9	5.5
	lexical	1.4	10.2	2.2	8.9
Croatian→Serbian		overall	*trebati*	overall	*trebati*
	ΣERR	6.3	21.5	7.8	20.3
	inflection	2.1	10.4	2.7	10.0
	order	0.3	0.9	0.2	1.2
	omission	0.8	1.6	0.3	1.7
	addition	0.4	3.0	0.2	2.9
	lexical	2.7	5.7	4.5	4.5

Table 4: Automatic evaluation scores BLEU and CHRF3 and classified edit operations on Serbian↔Croatian post-edited data.

5.1 Croatian-Serbian translation

The manual analysis revealed that for the Croatian-Serbian translation in both directions, constructions involving the verb *trebati* pose the most problems, both for the SMT as well as the RBMT system. Therefore, the segments containing this verb were isolated and analysed separately. The automatic evaluation scores are presented in Table 4, both for the whole test set as well as for the segments containing *trebati*.

- the overall performance is better for the SMT system, mainly due to less lexical errors;
- the RBMT system handles better the constructions with *trebati* producing less inflectional and lexical errors which are the predominant error types produced in these constructions
- both systems perform slightly better for translation into Croatian, but *trebati* constructions are better translated into Serbian by both systems. A probable reason for this is the different nature of the used test texts.

Especially problematic structures for both systems are long range dependencies where the main verb(s) is/are separated from the verb *trebati*. Furthermore, mistranslations were detected because of impersonal constructions and conditional forms, which are more problematic for the RBMT system. In addition, the meaning *to need* is often incorrectly translated by both systems, especially from Serbian into Croatian.

All in all, there is no significant difference between the performance of the SMT and the RBMT approach. Nevertheless, the systems do not always fail in the same way of the same segment, which indicates that a hybrid approach for this language pair could be beneficial.

5.2 Slovenian-to-Serbian translation

Manual evaluation has shown that the most frequent problems in Slovenian→Serbian post-edited translations are the future tense and the structures involving the Slovenian conjunction *tudi* (*also/too*). There-

Slovenian→Serbian	BLEU sl→sr	BLEU sl→hr→hr	CHRF3 sl→sr	CHRF3 sl→hr→sr
standard test	23.5	25.4	43.2	44.4
overall	70.3	71.7	81.6	82.8
future+*tudi*	67.8	73.8	78.3	84.0

Slovenian→Serbian	standard sl-sr	standard sl-hr-sr	post-edited sl-sr	post-edited sl-hr-sr	pe future+*tudi* sl-sr	pe future+*tudi* sl-hr-sr
ΣERR	62.8	62.1	14.3	15.8	16.8	15.2
inflection	3.5	3.6	2.3	2.1	3.6	2.3
order	4.0	4.9	0.8	3.2	1.4	1.9
omission	14.4	12.3	4.3	2.6	3.8	2.6
addition	9.1	9.6	1.1	1.6	0.9	2.3
lexical	31.8	31.8	5.9	6.2	7.1	6.2

Table 5: Automatic evaluation scores BLEU and CHRF3 and classified edit operations on Slovenian→Serbian post-edited data.

fore, sentences containing these two structures were identified and analysed separately.

For this translation direction, an additional preliminary experiment has been carried out, namely an attempt to improve the translation quality by two-stage (bridge, pivot) translation via Croatian.

Table 5 shows the overall post-edited results as well as the results on segments containing future tense and *tudi* both for direct as well as for two-stage SMT system. In addition, results for the overall standard test set are also shown for both systems. The following can be observed from the presented results:

- as expected, the automatic scores on post-edited test are lower than for Croatian-Serbian translation, but not so much lower as for the standard (suboptimal) test sets;
- scores on segments containing future tense and *tudi* are lower than overall scores;
- two-stage translation via Croatian generally helps, especially for the *problematic* segments – it reduces the number of inflectional edits, omissions and lexical edits;
- main disadvantage of two-stage translation is the increased amount of reordering errors.

The results show that two-stage translation has a good potential and should be further investigated. Further investigation should also include other translation directions from and into Slovenian, using different types of data sets.

6 Summary and outlook

This work represents a first step in the systematic evaluation of MT results between Croatian, Serbian and Slovenian and it has already shown several interesting results.

The analysis has revealed that the differences between the structural properties represent the most prominent issue for all translation directions. For translation between Croatian and Serbian, the constructions involving the verb *trebati* (*should/need*) definitely represent the larger obstacle for both translation directions and for both MT approaches, statistical as well as rule-based. However, the systems do not fail in the same way on the same segments, therefore hybrid systems should be investigated in future work.

For translations from Slovenian into Serbian, future tense represents one of the dominant issues followed by conjunction/adverb *tudi*. Other translation directions involving Slovenian have to be explored in future work. Two-stage translation via Croatian improves significantly the performance of the segments containing those problematic structures, the rest of the segments, however, are partially improved and partially deteriorated by introducing reordering errors and should be further investigated.

Future work should also include working on the identified issues, namely improving the systems by targeting the verb *trebati* and the Slovenian future tense. Also, other MT methods, such as hierarchical phrase-based and neural approach, should be investigated.

Acknowledgments

This work has emerged from research supported by TRAMOOC project (Translation for Massive Open Online Courses) partially funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under grant agreement number 644333 and by the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight). The research leading to these results has also received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and the Swiss National Science Foundation grant IZ74Z0_160501 (ReLDI).

References

Mihael Arčan, Maja Popović, and Paul Buitelaar. 2016. Asistent – a machine translation system for Slovene, Serbian and Croatian. In *Proceedings of the 10th Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia, September.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 412–418, Copenhagen.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, Reykjavik, Iceland, May.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. Collaborative development of a rule-based machine translator between Croatian and Serbian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. Washington DC.

David Kolovratník, Natalia Klyueva, and Ondřej Bojar. 2009. Statistical Machine Translation Between Related and Unrelated Languages. In *Proceedings of the Conference on Theory and Practice of Information Technologies (ITAT-09)*, Kralova Studna, Slovakia, September.

Vladislav Kubon and Jernej Vičič. 2014. A comparison of mt methods for closely related languages: a case study on czech - slovak language pair. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 92–98, Doha, Qatar, October. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Hrvoje Peradin, Filip Petkovski, and Francis Tyers. 2014. Shallow-transfer rule-based machine translation for the Western group of South Slavic languages. In *Proceedings of the 9th SaLTMiL Workshop on Free/open-Source Language Resources for the Machine Translation of Less-Resourced Languages*, pages 25–30, Reykjavik, Iceland, May.

Maja Popović and Mihael Arčan. 2015. Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In *18th Annual Conference of the European Association for Machine Translation (EAMT-15)*, Antalya, Turkey, May.

Maja Popović and Nikola Ljubešić. 2014. Exploring cross-language statistical machine translation for closely related South Slavic languages. In *Proceedings of the EMNLP14 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 76–84, Doha, Qatar, October.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, (96):59–68, October.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 392–395, Lisbon, Portugal, September.

Víctor Manuel Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 256–273, Lisbon, Portugal, September.

Jorg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.

Antonio Toral, Raphael Rubino, and Gema Ramírez-Sánchez. 2016. Re-assessing the Impact of SMT Techniques with Human Evaluation: a Case Study on English-Croatian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Francis M. Tyers and Murat Alperen. 2010. South-East European Times: A parallel corpus of the Balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Valetta, Malta, May.

Jernej Vičič. 2008. Rapid development of data for shallow transfer RBMT translation systems for highly inflective languages. In *Proceedings of the 6th Conference on Language Technologies*, Ljubljana, Slovenia, October.