Exploring cross-language statistical machine translation for closely related South Slavic languages

Maja Popović
DFKI Berlin, Germany
maja.popovic@dfki.de

Nikola Ljubešić
University of Zagreb, Croatia
nikola.ljubesic@ffzg.hr

Abstract

This work investigates the use of cross-language resources for statistical machine translation (SMT) between English and two closely related South Slavic languages, namely Croatian and Serbian. The goal is to explore the effects of translating from and into one language using an SMT system trained on another. For translation into English, a loss due to cross-translation is about 13% of BLEU and for the other translation direction about 15%. The performance decrease for both languages in both translation directions is mainly due to lexical divergences. Several language adaptation methods are explored, and it is shown that very simple lexical transformations already can yield a small improvement, and that the most promising adaptation method is using a Croatian-Serbian SMT system trained on a very small corpus.

1 Introduction

Statistical machine translation has become widely used over the last decade – open source tools such as Moses (Koehn et al., 2007) make it possible to build translation systems for any language pair within days, or even hours. However, the prerequisite is that appropriate bilingual training data is available, which is actually one of the most severe limitations of the statistical approach – large resources are only available for a few language pairs and domains. Therefore exploiting language closeness can be very convenient if there are no appropriate corpora containing the desired language, but it is possible to acquire corpora containing a closely related one. Croatian and Serbian are very close languages, and both[1] are under-resourced in terms of free/open-source language resources and tools, especially in terms of parallel bilingual corpora. On the other hand, Croatian has recently become the third official South Slavic language in the EU[2], and Serbian[3] is the official language of a candidate member state. Therefore investigating cross-language translation for these two languages can be considered very useful.

Both languages belong to the South-Western Slavic branch. As Slavic languages, they have a free word order and are highly inflected. Although they exhibit a large overlap in vocabulary and a strong morphosyntactic similarity so that the speakers can understand each other without difficulties, there is a number of small, but notable and frequently occurring differences between them.

In this paper, we investigate the impact of these differences on cross-language translation. The main questions are:

- How much will the translation performance decrease if a Serbian-English SMT system is used for translation from and into Croatian? (and the other way round)
- What are the possibilities for diminishing this performance decrease?

1.1 Related work

First publications dealing with statistical machine translation systems for Serbian-English (Popović et al., 2005) and for Croatian-English (Ljubešić et al., 2010) are reporting results of first steps on small bilingual corpora. Recent work on Croatian-English pair describes building a parallel corpus in the tourism domain by automatic web harvesting (Esplà-Gomis et al., 2014) and results of a SMT system built on this parallel corpus which yielded significant improvement (10%

[1] as well as other South Slavic languages

[2] together with Slovenian and Bulgarian

[3] together with Bosnian and Montenegrin

Language Technology for Closely Related Languages and Language Variants (LT4CloseLang), pages 76–84,
October 29, 2014, Doha, Qatar.

BLEU) over the Google baseline in the tourism domain (Toral et al., 2014). A rule-based Apertium system (Peradin et al., 2014) has been recently developed for translation from and into Slovenian (also closely related language, but more distant).

Techniques simpler than general SMT such as character-level translation have been investigated for translation between various close language pairs, where for the South Slavic group the Bulgarian-Macedonian pair has been explored (Nakov and Tiedemann, 2012). Character-based translation has also been used for translating between Bosnian and Macedonian in order to build pivot translation systems from and into English (Tiedemann, 2012).

Developing POS taggers and lemmatizers for Croatian and Serbian and using Croatian models on Serbian data has been explored in (Agić et al., 2013).

To the best of our knowledge, a systematic investigation of cross-language translation systems involving Croatian and Serbian, thereby exploiting benefits from the language closeness and analyzing problems induced by language differences has not been carried out yet.

2 Language characteristics

2.1 General characteristics

Croatian and Serbian, as Slavic languages, 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.

As for verbs, person and many tenses are expressed by the suffix, and the subject pronoun (e.g. I, we, it) is often omitted (similarly as in Spanish and Italian). In addition, negation of three quite important verbs, “biti” (to be, auxiliary verb for past tense, conditional and passive voice), “imati” (to have) and “ht(j)eti” (to want, auxiliary verb for the future tense), is formed by adding the negative particle to the verb as a prefix.

As for syntax, both languages have a quite free word order, and there are no articles.

2.2 Differences

The main differences between the languages are illustrated by examples in Table 1.

The largest differences between the two languages are in the vocabulary. Months have Slavic-derived names in Croatian whereas Serbian uses standard set of international Latin-derived names. A number of other words are also completely different (1), and a lot of words differ only by one or two letters (2). In addition, Croatian language does not transcribe foreign names and words, whereas phonetical transcriptions are usual in Serbian although original writing is allowed too (3).

Apart from lexical differences, there are also structural differences mainly concerning verbs. After modal verbs such as “morati” (to have to) or “moći” (can) (4), the infinitive is prescribed in Croatian (“moram raditi”), whereas the construction with particle “da” (that/to) and present tense (“moram da radim”) is preferred in Serbian. An inspection of the Croatian and Serbian web corpora[4] (Ljubešić and Klubička., 2014) shows the prescription being followed by identifying 1286 vs. 29 occurrences of the two phrases in the Croatian and 40 vs. 322 occurrences in the Serbian corpus. It is important to note that the queried corpora consist of texts from the Croatian and Serbian top-level web domain and that the results in discriminating between Croatian and Serbian language applied to these corpora are not used at this point.

The mentioned difference partly extends to the future tense (5), which is formed in a similar manner to English, using present of the verb "ht(j)eti" as 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”), whereas in Croatian only the infinitive form is used (“ja ću raditi”). This is, again, followed by corpus evidence with 0 vs. 71 occurrences of the phrases in the Croatian corpus and 13 vs. 22 occurrences in the Serbian corpus. Another difference regarding future tense exists when the the auxiliary and main verb are reversed (5b): 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”).

[4]the corpora can be queried via `http://nl.ijs.si/noske/`

		Croatian	Serbian	English
vocabulary				
1)	word level	gospodarstvo	ekonomija	economy
		tjedan	nedelja	week
		tisuća	hiljada	one thousand
	months	siječanj	januar	January
2)	character level	točno	tačno	accurate
		Europa	Evropa	Europe
		vjerojatno	verovatno	probably
		vijesti	vesti	news
		terorist	terorista	terrorist
3)	transcription	Washington	Vašington	Washington
structure (verbs)				
4)	modal verbs	moram raditi	moram da radim	I have to work
		mogu raditi	mogu da radim	I can work
5)	future tense a)	ja ću raditi	ja ću da radim	I will work
	b)	radit ću	radiću	I will work
6)	"trebati" = should a)	trebam raditi	treba da radim	I should work
		trebaš raditi	treba da radiš	you should work
	= need b)	trebam posao	treba mi posao	I need a job
		Petar treba knjige	Petru trebaju knjige	Petar needs books

Table 1: Examples of main differences between Croatian and Serbian.

Corpus evidence follows this as well with 611 vs. 9 occurrences in the Croatian corpus and 4 vs. 103 occurrences in the Serbian one. A very important difference concerns the verb "trebati" (to need, should) (6). In Croatian, the verb takes the tense according to the subject and it is transitive as in English. In Serbian, when it means "should" (6a) it is impersonal followed by "da" and the present of the main verb ("treba da radim"). When it means "to need" (6b), the verb is conjugated according to the needed object ("treba" (job), "trebaju" (books)), and the subject which needs something (I, Petar) is an indirect grammatical object in dative case ("meni", "Petru").

Apart from the described differences, there is also a difference in scripts: Croatian uses only the Latin alphabet whereas Serbian uses both Latin and Cyrillic scripts[5]. However, this poses no problem regarding corpora because a Cyrillic Serbian text can be easily transliterated into Latin.

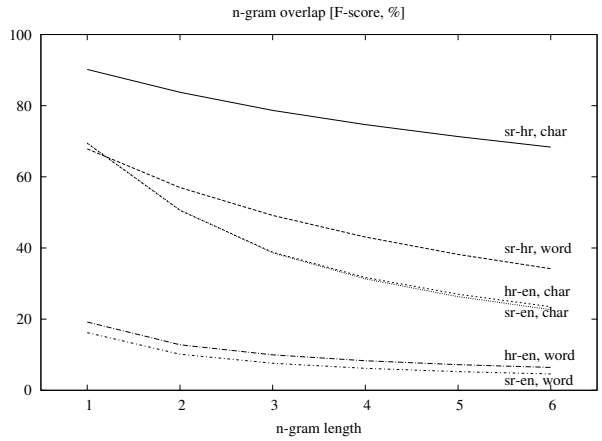

Figure 1: n-gram overlap on word level and on character level between Croatian-Serbian, Croatian-English and Serbian-English.

The idea of Figure 1 is to illustrate the closeness and the differences between the two close languages of interest by numbers: overlapping of

[5]During the compilation process of the Serbian web corpus (Ljubešić and Klubička., 2014), 16.7% of retrieved text was written in the Cyrillic script.

word level and character level n-grams for $n = 1, ...6$ in training, development and test corpora together is presented via the F-score. In order to give a better insight, overlaps with English are calculated as well. It can be seen that the Croatian-Serbian overlap on character level is very high, and still rather high on the word level. Character overlaps with English are below the Croatian-Serbian overlap on the word level, whereas the word level overlaps with English are very low.

3 Translation experiments

In order to explore effects of the described language differences on cross-language SMT, four translation systems have been built: Croatian→English, Serbian→English, English→Croatian and English→Serbian. For the sake of brevity and clarity, we will use the terms "corresponding source/output" when the test language is same as the language used for training, and "other source/output" when the cross-language translation is performed. For translation into English, the translation outputs of the other source text and its adapted variants are compared to the translation output of the corresponding source test with respect to the English reference. For translation from English, the other translation output and its adapted versions are compared to the corresponding output with respect to the corresponding reference. The investigated adaptation methods are described in the next section.

3.1 Language adaptation methods

The following methods were investigated for adaptation of the test set in the other language:

- lexical conversion of the most frequent words (*conv*);

  The most frequent[6] different words together with simple morphological variations are replaced by the words in the corresponding language. This method is simple and fast, however it is very basic and also requires knowledge of the involved languages to be set up. It can be seen as a very first step towards the use of a rule-based Croatian-Serbian system.

- Croatian-Serbian translation system trained on three thousand parallel sentences (*3k*);

  This method does not require any language knowledge, and a small bilingual corpus is often not very difficult to acquire. It is even not very difficult to create it manually from a monolingual corpus by translating it, although in that case the language knowledge is needed.

- Croatian-Serbian translation system trained on the large parallel corpus (*200k*);

  This method is interesting in order to see the upper limits of the adaptation, however it is not realistic – if a large in-domain corpus is available in both languages, there is no need for cross-language translation, but pivoting or synthetic corpora can be used.

The language adaptation is performed in the following way: for translation into English, the other language test set is first preprocesssed, i.e. converted or translated into the corresponding language, and then translated. For the other translation direction, the English test is translated into the other language and then converted/translated into the corresponding one.

In addition, training a system using the converted corpus has also been investigated for all translation directions.

4 Experimental set-up

The enhanced version[7] of the SEtimes corpus (Tyers and Alperen, 2010) is used for translation experiments. The corpus is based on the content published on the SETimes.com news portal which publishes "news and views from Southeast Europe" in ten languages: Bulgarian, Bosnian, Greek, English, Croatian, Macedonian, Romanian, Albanian and Serbian. We used the parallel trilingual Croatian-English-Serbian part of the corpus. The detailed corpus statistic is shown in Table 2. The Croatian language is further referred to as *hr*, Serbian as *sr* and English as *en*.

The translation system used is the phrase-based Moses system (Koehn et al., 2007). The evaluation metrics used for assessment of the translations are the BLEU score (Papineni et al., 2002) and the F-score, which also takes recall into account and generally better correlates with human rankings which has been shown in (Melamed et al., 2003) and confirmed in (Popović, 2011). For

[6] occurring $\geq$ 1000 times in the training corpus

[7] http://nlp.ffzg.hr/resources/corpora/setimes/

		Croatian (hr)	Serbian (sr)	English (en)
Train	sentences		197575	
	avg sent length	22.3	22.5	23.9
	running words	4410721	4453579	4731746
	vocabulary	149416	144544	76242
Dev	sentences		995	
	avg sent length	22.2	22.5	24.0
	running words	22125	22343	23896
	running OOVs	1.7%	1.6%	0.8%
Test	sentences		1000	
	avg sent length	22.3	22.4	23.8
	running words	22346	22428	23825
	running OOVs	1.5%	1.4%	0.7%

Table 2: Corpus statistics

translation into Croatian and Serbian, F-scores on character level are also calculated.

5 Results

5.1 Croatian↔Serbian language adaptation

This section presents the results of conversion and translation between Croatian and Serbian in order to better understand advantages and disadvantages of each of the adaptation methods. The effects of each method on translation into and from English will be reported in the next section.

Table 3 shows the BLEU and F-scores as well as the percentage of running OOVs for each adaptation method. If no adaptation is performed (first row), the word level scores are about 40%, CHARF score is close to 75% , and a large number of OOVs is present – 13% of running words are unseen. A large portion of these words differ only by one or two characters, and for a standard SMT system there is no difference between such words and completely distinct ones.

The *conv* method, i.e. simple replacement of a set of words, already makes the text more close: it reduces the number of OOVs by 3-5% and improves the scores by 3%. The best results are obtained, as it can be expected, by *200k* adaptation, i.e. translation using the large Croatian-Serbian training corpus; the amount of OOVs in the adapted text is comparable with the text in the corresponding language (presented in Table 2). The *3k* translation system, being the most suitable for "real-word" tasks and improving significantly the text in the other language (almost 10% reduction of OOVs and 13% increase of scores) seems to be the most promising adaptation method.

5.2 Croatian/Serbian↔English translation

The translation results into and from English are presented in Table 4. It can be seen that the BLEU/WORDF loss induced by cross-language translation is about 12-13% for translation into English and about 13-15% for the other direction. The effects of language adaptation methods are similar for all translation directions: the simple lexical conversion *conv* slightly improves the translation outputs, and the best option is to use the *200k* translation system. The small training corpus achieves, of course, less improvement than the large corpus. On the other hand, taking into account the significant improvement over the original of the text of the other language (about 9%) and the advantages of the method discussed in Sections 3.1 and 5.1, this performance difference is actually not too large. Future work should explore techniques for improvement of such systems.

Last two rows in each table represent the results of the additional experiment, namely using the converted other language corpus for training. However, the results do not outperform those obtained by (much faster) conversion of the source/output, meaning that there is no need for retraining the translation system – it is sufficient to adapt only the test source/output.

Translation examples

Table 5 presents two translation examples: the source/reference sentence in all three languages, the cross-language translation output, the trans-

direction	method	BLEU	WORDF	CHARF	OOV
	none	40.1	43.1	74.7	13.3
hr→sr	conv	43.7	46.3	76.4	10.7
	3k	54.8	55.9	80.8	4.6
	200k	64.3	65.4	85.2	1.4
sr→hr	conv	43.5	46.1	76.3	8.5
	3k	54.0	55.9	80.9	4.3
	200k	64.1	65.3	85.1	1.4

Table 3: BLEU and F-scores for Croatian-Serbian conversion and translation used for adaptation.

lation outputs of adapted sources, as well as the translation output of the corresponding source. The examples are given only for translation into English, and the effects for the other translation direction can be observed implicitly. Generally, the main source of errors are OOV words, but structural differences also cause problems.

For the first sentence (1), the *conv* method is sufficient for obtaining a perfect cross-translation output: the obstacles are three OOV words, all of them being frequent and thus converted. The outputs obtained by *3k* and *200k* methods as well as the output for the corresponding language are exactly the same and therefore not presented.

The second sentence (2) is more complex: it contains three OOV words, two of which are not frequent and thus not adapted by *conv*, and one future tense i.e. a structural difference. The OOV words do not only generate lexical errors (untranslated words) but also incorrect word order ("from 17 dječjih kazališta"). The *conv* method is able to repair only the month name, whereas other errors induced by language differences[8] are still present. The *3k* translation system resolves one more OOV word ("theater") together with its position, as well as the future tense problem, but the third OOV word "children's" is still untranslated and in the wrong position. This error is fixed only when *200k* translation system is used, since the word occurs in the large corpus but not in the small one. It should be noted that the word is, though, an OOV only due to the one single letter and probably could be dealt with by character-based techniques (Nakov and Tiedemann, 2012) which should be investigated in future work.

[8]It should be noted that errors not related to the language differences are out of the scope of this work.

6 Conclusions

In this work, we have examined the possibilities for using a statistical machine translation system built on one language and English for translation from and into another closely related language. Our experiments on Croatian and Serbian showed that the loss by cross-translation is about 13% of BLEU for translation into English and 15% for translation from English.

We have systematically investigated several methods for language adaptation. It is shown that even a simple lexical conversion of limited number of words yields improvements of about 2% BLEU, and the Croatian-Serbian translation system trained on three thousand sentences yields a large improvement of about 6-9%. The best results are obtained when the translation system built on the large corpus is used; however, it should be taken into account that such scenario is not realistic.

We believe that the use of a small parallel corpus is a very promising method for language adaptation and that the future work should concentrate in improving such systems, for example by character-based techniques. We also believe that a rule-based Croatian-Serbian system could be useful for adaptation, since the translation performance has been improved already by applying a very simple lexical transfer rule. Both approaches will be investigated in the framework of the ABUMATRAN project[9].

Depending on the availability of resources and tools, we plan to examine texts in other related languages such as Slovenian, Macedonian and Bulgarian (the last already being part of ongoing work in the framework of the QTLEAP project[10]), and also to do further investigations on the Croatian-Serbian language pair.

[9]http://abumatran.eu/

[10]http://qtleap.eu/

(a) translation into English

training	source	BLEU	WORDF
sr→en	hr	29.8	34.1
	hr-sr.conv	32.3	36.4
	hr-sr.3k	37.6	41.1
	hr-sr.200k	42.3	45.6
	sr	42.9	46.0
hr→en	sr	31.4	35.5
	sr-hr.conv	32.8	36.8
	sr-hr.3k	37.2	40.8
	sr-hr.200k	41.7	44.9
	hr	43.2	46.3
sr-hr.conv→en	hr	32.2	36.2
hr-sr.conv→en	sr	33.5	37.4

(b) translation from English

reference	output	BLEU	WORDF	CHARF
hr	sr	20.6	25.4	62.7
	sr-hr.conv	22.8	27.4	64.2
	sr-hr.3k	29.3	33.4	68.5
	sr-hr.200k	33.5	37.2	71.2
	hr	35.5	38.9	72.1
sr	hr	20.3	25.3	62.7
	hr-sr.conv	22.6	27.4	64.2
	hr-sr.3k	29.8	33.7	68.4
	hr-sr.200k	34.0	37.5	71.3
	sr	35.3	38.5	72.1
sr	en→hr-sr.conv	22.6	27.4	64.2
hr	en→sr-hr.conv	23.2	27.7	64.2

Table 4: BLEU, WORDF and CHARF scores for translation (a) into English; (b) from English.

Acknowledgments

This work has been supported by the QTLEAP project – EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLEAP: Quality Translation by Deep Language Engineering Approaches" and the ABU-MATRAN project – EC's FP7 (FP7/2007-2013) under grant agreement number PIAP-GA-2012-324414: "ABU-MATRAN: Automatic building of Machine Translation".

References

Željko Agić, Nikola Ljubešić and Danijela Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian, In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August.

Miquel Esplà-Gomis and Filip Klubička and Nikola Ljubešić and Sergio Ortiz-Rojas and Vassilis Papavassiliou and Prokopis Prokopidis 2014. Comparing Two Acquisition Systems for Automatically Building an English-Croatian Parallel Corpus from Multilingual Websites, In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, May.

Nikola Ljubešić and Filip Klubička 2014. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian, In Proceedings of the 9th Web as Corpus Workshop (WaC-9), Gothenburg, Sweden, April.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran,

	sr	Pregled poslovnih i **ekonomskih vesti** sa Balkana od 15. **avgusta**.
1)	hr	Pregled poslovnih i **gospodarskih vijesti** s Balkana od 15. **kolovoza**.
	en	A review of business and economic news from the Balkans since 15 **August**.
training:	sr→en	
source:	hr	A review of business and **gospodarskih vijesti** from the Balkans since 15 **kolovoza** .
	hr-sr.conv	A review of business and economic news from the Balkans since 15 August .

	sr	Srpski grad Subotica *biće* domaćin 16. izdanja Međunarodnog festivala **dečjih pozorišta** od 17. do 23. **maja**.
2)	hr	Subotica u Srbiji *bit će* domaćin 16 . Međunarodnog festivala **dječjih kazališta** od 17. do 23. **svibnja**.
	en	Subotica, Serbia, *will host* the 16th edition of the International Festival of **Children's Theatres** from **May** 17th to **May** 23rd .
training:	sr→en	
source:	hr	Subotica in Serbia *will be will host* the 16th International Festival from 17 **dječjih kazališta** to 23 **svibnja**.
	hr-sr.conv	Subotica in Serbia *will be will host* the 16th International Festival from 17 **dječjih kazališta** to 23 May.
	hr-sr.3k	Subotica in Serbia will host the 16th International Theatre Festival from 17 **dječjih** to 23 May.
	hr-sr.200k	Subotica in Serbia will host the 16th International Children's Theatre Festival from 17 to 23 May.
	sr	The Serbian town of Subotica will host the 16th edition of the International Children's Theatre Festival from 17 to 23 May.

Table 5: Two examples of cross-translation of Croatian source sentence into English using Serbian→English translation system.

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*, pages 177–180, Prague, Czech Republic, June.

Nikola Ljubešić, Petra Bago and Damir Boras. 2010. Statistical machine translation of Croatian weather forecast: How much data do we need?, In *Proceedings of the ITI 2010 32nd International Conference on Information Technology Interfaces*, pages 91–96, Cavtat, Croatia, June.

I.Dan Melamed, Ryan Green and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 61–63, Edmonton, Canada, May/June.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages, In *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 301–305, Jeju, Republic of Korea, July.

Kishore Papineni, Salim Roukos, Todd Ward and Wie-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

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, David Vilar, Hermann Ney, Slobodan Jovičić and Zoran Šarić. 2005. Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian–English Statistical Machine Translation In *Proceedings of the ACL-05 Workshop on Building and Using Parallel*

Texts: Data-Driven Machine Translation and Beyond, pages 119–124, Ann Arbor, MI, June.

Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics, In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 104–107, Edinburgh, Scotland, July.

Jörg Tiedemann. 2012. Character-based pivot translation for under-resourced languages and domains, In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 141–151, Avignon, France, April.

Antonio Toral, Raphael Rubino, Miquel Esplà-Gomis, Tommi Pirinen, Andy Way and Gema Ramirez-Sanchez 2014. Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain, In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, pages 221–224, Dubrovnik, Croatia, June.

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.