Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages

Maja Popović
DFKI – Language Technology Lab
Berlin, Germany
maja.popovic@dfki.de

Mihael Arčan
Insight Centre for Data Analytics
National University of Galway, Ireland
mihael.arcan@insight-centre.org

Abstract

The best way to improve a statistical machine translation system is to identify concrete problems causing translation errors and address them. Many of these problems are related to the characteristics of the involved languages and differences between them. This work explores the main obstacles for statistical machine translation systems involving two morphologically rich and under-resourced languages, namely Serbian and Slovenian. Systems are trained for translations from and into English and German using parallel texts from different domains, including both written and spoken language. It is shown that for all translation directions structural properties concerning multi-noun collocations and exact phrase boundaries are the most difficult for the systems, followed by negation, preposition and local word order differences. For translation into English and German, articles and pronouns are the most problematic, as well as disambiguation of certain frequent functional words. For translation into Serbian and Slovenian, cases and verb inflections are most difficult. In addition, local word order involving verbs is often incorrect and verb parts are often missing, especially when translating from German.

1 Introduction

The statistical approach to machine translation (SMT), in particular phrase-based SMT, has become widely used in the last years: open source tools such as Moses (Koehn et al., 2007) have made it possible to build translation systems for any language pair, domain or text type within days. Despite the fact that for certain language pairs, e.g. English-French, high quality SMT systems have been developed, a large number of languages and language pairs have not been (systematically) investigated. The largest study about European languages in the Digital Age, the META-NET Language White Paper Series[1] in year 2012 showed that only English has good machine translation support, followed by moderately supported French and Spanish. More languages are only fragmentary supported (such as German), whereby the majority of languages are weakly supported. Many of those languages are also morphologically rich, which makes the SMT task more complex, especially if they are the target language. A large part of weakly supported languages consists of Slavic languages, where both Serbian and Slovenian belong. Both languages are part of to the South Slavic language branch, Slovenian[2] being the third official South Slavic language in the EU and Serbian[3] is the official language of a candidate member state. For all these reasons, a systematic investigation of SMT systems involving these two languages and defining the most important errors and problems can be very very beneficial for further studies.

In the last decade, several SMT systems have been built for various South Slavic languages and English, and for some systems certain morpho-syntactic transformations have been applied more

[1] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison
[2] together with Croatian and Bulgarian
[3] together with Bosnian and Montenegrin

or less successfully. However, all the experiments are rather scattered and no systematic or focused research has been carried out in order to define actual translation errors as well as their causes.

This paper reports results of an extensive error analysis for four language pairs: Serbian and Slovenian with English as well as with German, which is also a challenging language – complex (both as a source and as a target language) and still not widely supported. SMT systems have been built for all translation directions using publicly available parallel texts originating from several different domains including both written and spoken language.

2 Related work

One of the first publications dealing with SMT systems for Serbian-English (Popović et al., 2005) and Slovenian-English (Sepesy Maučec et al., 2006) are reporting first results using small bilingual corpora. Improvements for translation into English are also reported by reducing morpho-syntactic information in the morphologically rich source language. Using morpho-syntactic knowledge for the Slovenian-English language pair was shown to be useful for both translation directions in (Žganec Gros and Gruden, 2007). However, no analysis of results has been carried out in terms of what were actual problems caused by the rich morphology and which of those were solved by the morphological preprocessing.

Through the transLectures project,[4] the Slovenian-English language pair became available in the 2013 evaluation campaign of IWSLT.[5] They report the BLEU scores of TED talks translated by several systems, however a deeper error analysis is missing (Cettolo et al., 2013).

Recent work in SMT also deals with the Croatian language, which is very closely related to Serbian. First results for Croatian-English are reported in (Ljubešić et al., 2010) on a small weather forecast corpus, and an SMT system for the tourist domain is presented in (Toral et al., 2014). Furthermore, SMT systems for both Serbian and Croatian are described in (Popović and Ljubešić, 2014), however only translation errors caused by language mixing are analysed, not the problems related to the languages themselves.

[4] https://www.translectures.eu/

[5] International Workshop on Spoken Language Translation, http://workshop2013.iwslt.org/

Different SMT systems for subtitles were developed in the framework of the SUMAT project,[6] including Serbian and Slovenian (Etchegoyhen et al., 2014). However, only the translations between them have been carried out as an example of closely related and highly inflected languages.

3 Language characteristics

Serbian (referred to as "sr") and Slovenian ("sl"), as Slavic languages, have quite free word order and are highly inflected. The inflectional morphology is very rich 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, 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* (sr/sl)" (to be), "*imati* (sr) / *imeti* (sl)" (to have) and "*hteti* (sr) / *hoteti* (sl)" (to want), is formed by adding the negative particle to the verb as a prefix. In addition, there are two verb aspects, namely many verbs have perfective and imperfective form(s) depending on the duration of the described action. These forms are either different although very similar or are distinguished only by prefix.

As for syntax, both languages have a quite free word order, and there are no articles, neither indefinite nor definite.

Although the two languages share a large degree of morpho-syntactic properties and mutual intelligibility, a speaker of one language might have difficulties with the other. The language differences are both lexical (including a number of false friends) as well as grammatical (such as local word order, verb mood and/or tense formation, question structure, dual in Slovenian, usage of some cases).

4 SMT systems

In order to systematically explore SMT issues related to the targeted languages, five different domains were used in total. However, not all do-

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

(a) number of sentences

# of Sentences	sl-en	sl-de	sr-en	sr-de
DGT	3.2M	3M	/	/
Europarl	600k	500k	/	/
EMEA	1M	1M	/	/
OpenSubtitles	1.8M	1.8M	1.8M	1.8M
SEtimes	/	/	200k	/

(b) average sentence length

Avg. Sent. Length	sl	sr	en	de
DGT	16.0	/	17.3	16.6
Europarl	23.4	/	27.0	25.4
EMEA	12.7	/	12.3	11.8
OpenSubtitles	7.7	7.6	9.2	8.9
SEtimes	/	22.4	23.8	/

Table 1: Corpora characteristics.

mains were used for all language pairs due to unavailability. It should be noted that according to the META-NET White Papers, both languages have minimal support, with only fragmentary text and speech resources. For the Slovenian-English and Slovenian-German language pairs, four domains were investigated: DGT translation memories provided by the JRC (Steinberger et al., 2012), Europarl (Koehn, 2005), European Medicines Agency corpus (EMEA) in the pharmaceutical domain, as well as the OpenSubtitles[7] corpus. All the corpora are downloaded from the OPUS web site[8] (Tiedemann, 2012). For the Serbian language, only two domains were available: the enhanced version of the SEtimes corpus[9] (Tyers and Alperen, 2010) containing "news and views from South-East Europe" for Serbian-English, and OpenSubtitles for the Serbian-English and Serbian-German language pairs. It should be noted that all the corpora contain written texts except OpenSubtitles, which contains transcriptions and translations of spoken language thus being slightly peculiar for machine translation. On the other hand, this is the only corpus containing all language pairs of interest.

Table 1 shows the amount of parallel sentences for each language pair and domain (a) as well as the average sentence length for each language and domain (b). For each domain, a separate system has been trained and tuned on an unseen portion of in-domain data. Since the sentences in OpenSubtitles are significantly shorter than in other texts, the tuning and test sets for this domain contain 3000 sentences whereas all other sets contain 1000 sentences. Another remark regarding the OpenSubtitles corpus is that we trained our systems only on those sentence pairs, which were available in English as well as in German in order to have a completely same condition for all systems.

All systems have been trained using phrase-based Moses (Koehn et al., 2007), where the word alignments were build with GIZA++ (Och and Ney, 2003). The 5-gram language model was build with the SRILM toolkit (Stolcke, 2002).

5 Evaluation and error analysis

The evaluation has been carried out in three steps: first, the BLEU scores were calculated for each of the systems. Then, the automatic error classification has been applied in order to estimate actual translation errors. After that, manual inspection of language related phenomena leading to particular errors is carried out in order to define the most important issues which should be addressed for building better systems and/or develop better models.

5.1 BLEU scores

As a first evaluation step, the BLEU scores (Papineni et al., 2002) have been calculated for each of the translation outputs in order to get a rough idea about the performance for different domains and translation directions.

The scores are presented in Table 2:

- the highest scores are obtained for translations into English;
- the scores for translations into German are similar to those for translations into Slovenian and Serbian;
- the scores for Serbian and Slovenian are better when translated from English than when translated from German;
- the best scores are obtained for DGT (which contains a large number of repetitions), followed by EMEA (which is very specific domain); the worst scores are obtained for spoken language OpenSubtitles texts.

[7] http://www.opensubtitles.org/
[8] http://opus.lingfil.uu.se/
[9] http://nlp.ffzg.hr/resources/corpora/setimes/

Domain/Lang. pair	sl-en	sr-en	sl-de	sr-de	en-sl	de-sl	en-sr	de-sr
DGT	77.3	/	59.3	/	72.1	58.6	/	/
Europarl	58.9	/	33.8	/	56.0	36.5	/	/
EMEA	69.7	/	53.8	/	66.0	56.2	/	/
OpenSubtitles	38.4	33.2	21.5	22.4	26.2	19.6	22.8	18.4
SEtimes	/	43.8	/	/	/	/	35.8	/

Table 2: BLEU scores for all translation outputs.

In addition, all the BLEU scores are compared with those of Google Translate[10] outputs of all tests. All systems built in this work outperform the Google translation system by absolute difference ranges from 1 to 10%, confirming that the languages are weakly supported for machine translation.

5.2 Automatic error classification

Automatic error classification has been performed using Hjerson (Popović, 2011) and the error distributions are presented in Figure 1. For the sake of brevity and clarity, as well as for avoiding redundancies, the error distributions are not presented for all translation outputs, but have been averaged in the following way: since no important differences were observed neither between domains (except that OpenSubtitles translations exhibit more inflectional errors than others) nor between Serbian and Slovenian (neither as source nor as the target language), the errors are averaged over domains and two languages called "x". Thus, four error distributions are shown: translation from and into English, and translation from and into German.

The following can be observed:

- translations into English are "the easiest", mostly due to the small number of morphological errors; however, the English translation outputs contain more word order errors than Serbian and Slovenian ones;
- all error rates are higher in German translations than in English ones, but the mistranslation rate is especially high;
- German translation outputs have less morphological errors than Serbian and Slovenian translations from German; on the other hand, more reordering errors can be observed in German outputs;
- all error rates are higher in translations from German than from English, except inflections.

The results of the automatic error analysis are valuable and already indicate some promising directions for improving the systems, such as word order treatment and handling morphologic generation. Nevertheless, more improvement could be obtained if more precise guidance about problems and obstacles related to the language properties and differences were available (apart from the general ones already partly investigated in related work).

5.3 Identifying linguistic related issues

Automatic error analysis has already shown that that different language combinations show different error distributions. This often relates to linguistic characteristics of involved languages as well as to divergences between them. In order to explore those relations, manual inspection of about 200 sentences from each domain and language pair annotated by Hjerson together with their corresponding source sentences has been carried out.

As the result of this analysis, the following has been discovered:

- there is a number of frequent error patterns, i.e. obstacles (issues) for SMT systems
- nature and frequency of many issues depend on language combination and translation direction
- some of translation errors depend on the domain and text type, mostly differing for written and spoken language
- issues concerning Slovenian and Serbian both as source and as target languages are almost

[10] http://translate.google.com, performed on 27th February 2015

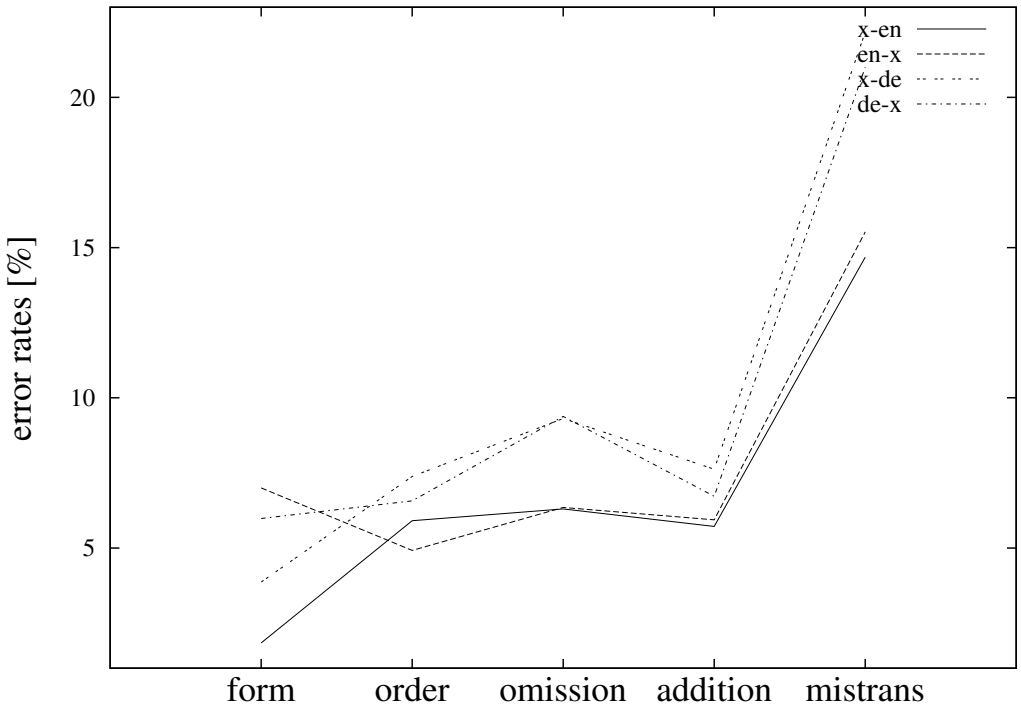

Figure 1: Error rates for five error classes: word form, word order, omission, addition and mistranslation; each error rate represents the percentage of particular (word-level) error type normalised over the total number of words.

the same – there are only few language specific phenomena.

The most frequent general issues[11], i.e. relevant for all translation directions, are:

- **noun collocations** (written language)

  Multi-word expressions consisting of a head noun and additional nouns and adjectives in English poses large problems for all translation directions, especially from and into English. Their formation is different in other languages and often the boundaries are not well detected so that they are translated to unintelligible phrases.

source	12th ”Tirana Fall” art and music festival
output*	12th ”Tirana collection fall of art and a music festival
reference	the ratings agency’s first Emerging Europe Sensitivity Index (EESI)
output	the first Index sensitivity Europe in the development of (EESI) this agency

- **negation**

  Due to the distinct negation forming, mainly concerning multiple negations in Serbian and Slovenian, negation structures are translated incorrectly.

reference	the prosecution has done nothing so far
source*	the prosecution has not done nothing so far
output	the prosecution is not so far had done nothing
source	Being a rector does not give someone the freedom
reference*	Being a rector does not give nobody the freedom
output*	Being a rector does not give some freedom

- **phrase boundaries and order**

  Phrase boundaries are not detected properly so that the group(s) of words are misplaced, often accompanied with morphological errors and mistranslations.

[11] Non-English parts of examples are literally translated into English and marked with *.

reference	of which about a fifth is used for wheat production
output	of which used to produce about one fifth of wheat
reference	But why have I brought this thing here?
output	This thing, but why am I here?
reference	The US ambassador to Sofia said on Wednesday .
output	Said on Wednesday , US ambassador to Sofia .

- **prepositions**

  Prepositions are mostly mistranslated, sometimes also omitted or added.

- **word order**

  Local word permutations mostly affecting verbs and surrounding auxiliary verbs, pronouns, nouns and adverbs.

Some of the frequent issues are strongly dependent on translation direction. For translation into English and German the issues of interest are:

- **articles**

  Due to the absence of articles in Slavic languages, a number of articles in English and German translation outputs are missing, and a certain number is mistranslated or added.

- **pronouns**

  The source languages are pro-drop, therefore a number of pronouns is missing in the English and German translation outputs.

- **possessive pronoun "*svoj*"**

  This possessive pronoun can be used for all persons ("my", "your", "her", "his", "our", "their") and it is difficult to disambiguate.

- **verb tense**

  Due to different usage of verb tenses in some cases, the wrong verb tense from the source language is preserved, or sometimes mistranslated. The problem is more prominent for translation into English.

- **agreement** (German target)

  A number of cases and gender agreements in German output is incorrect.

- **missing verbs** (German target)

  Verbs or verb parts are missing in German output, especially when they are supposed to appear at the end of the clause.

- **conjunction "*i*" (and)** (Serbian source)

  The main meaning of this conjunction is "and", but another meaning "also, too, as well" is often used too; however, it is usually translated as "and".

- **adverb "*lahko*"** (Slovenian source)

  This word is used for Slovenian conditional phrases which correspond to English constructions with modal verbs "can", "might", "shall", or adverbs "possibly"; the entire clause is often not translated correctly due to incorrect disambiguation.

For translation into Serbian and Slovenian, the most important obstacles are:

- **case**

  Incorrect case is used, often in combination with incorrect singular/plural and/or gender agreement.

- **verb inflection**

  Verb inflection does not correspond to the person and/or the tense; a number of past participles also has incorrect singular/plurar and/or gender agreement.

- **missing verbs**

  Verb or verb parts are missing, especially for constructions using auxiliary and/or modal verbs. The problem is more frequent when translating from German.

- **question structure** (spoken language)

  Question structure is incorrect; the problem is more frequent in Serbian where additional particles ("li" and "da li") should be used but are often omitted.

- **conjunction "*a*"** (Serbian target)

  This conjunction does not exist in other languages, it can be translated as "and" or "but", and its exact meaning is something in between. Therefore it is difficult to disambiguate the corresponding source conjunction.

- **"*-ing*" forms** (English source)

  English present continuous tense does not exist in other languages, and in addition, it is often difficult to determine if the word with the suffix "-ing" is a verb or a noun. Therefore words with the "-ing" suffix are difficult for machine translation.

- **noun-verb disambiguation** (English source)

  Apart from the words ending with the suffix "-ing", there is a number of other English words which can be used both as a noun as well as a verb, such as "offer", "search", etc.

- **modal verb "*sollen*"** (German source)

  This German modal verb can have different meanings, such as "should", "might" and "will" which is often difficult to disambiguate.

It has to be noted that some of the linguistic phenomena known to be difficult are not listed – the reason is that their overall number of occurrences in the analysed texts is low and therefore the number of related errors too. For example, German compounds are well known for posing problems to natural language processing tasks among which is machine translation – however, in the given texts only a few errors related to compounds were observed, as well as a low total number of compounds. Another similar case is the verb aspect in Serbian and Slovenian – some related errors were detected, but their count as well as the overall count of such verbs in the data is very small.

Therefore the structure and nature of the texts for a concrete task should always be taken into account. For example, for improvements of spoken language translation more effort should be put in question treatment than in noun collocation, and in technical texts the compound problem would probably be significant.

6 Conclusions and future work

In this work, we have examined several SMT systems involving two morphologically rich and under-resourced languages in order to identify the most important language related issues which should be dealt with in order to build better systems and models. The BLEU scores are reported as a first evaluation step, followed by automatic error classification which has captured interesting language related patterns in distributions of error types. The main part of the evaluation consisted of (manual) analysis of errors taking into account linguistic properties of both target and source language. This analytic analysis has defined a set of general issues which are causing errors for all translation directions, as well as sets of language dependent issues. Although many of these issues are already known to be difficult, they can be addressed only with the precise identification of concrete examples.

The main general issues are shown to be structural properties concerning multi-noun collocations and exact phrase boundaries, followed by negation formation, wrong, missing or added preposition as well as local word order differences. For translation into English and German, article and pronoun omissions are the most problematic, as well as disambiguation of certain frequent functional words. For translation into Serbian and Slovenian, cases and verb inflections are most difficult to handle. In addition, other problems concerning verbs are frequent as well, such as local word order involving verbs and missing verb parts (which is especially difficult when translating from German).

In future work we plan to address some of the presented issues practically and analyse the effects. An important thing concerning system improvement is that although most of the described issues are common for various domains, the exact nature of the texts desired for the task at hand should always be kept in mind. Analysis of issues for domains and text types not covered by this paper should be part of future work too.

Acknowledgments

This publication has emanated from research supported by the QT21 project – European Union's Horizon 2020 research and innovation programme under grant number 645452 as well as a research grant from Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289. We are grateful to the anonymous reviewers for their valuable feedback.

References

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December.

Etchegoyhen, Thierry, 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.

Koehn, Philipp, 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.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, September.

Ljubešić, Nikola, Petra Bago, and Damir Boras. 2010. Statistical machine translation of Croatian weather forecast: How much data do we need? In Lužar-Stiffler, Vesna, Iva Jarec, and Zoran Bekić, editors, *Proceedings of the ITI 2010 32nd International Conference on Information Technology Interfaces*, pages 91–96, Zagreb. SRCE University Computing Centre.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. pages 311–318, Philadelphia, PA, July.

Popović, Maja 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.

Popović, Maja, 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. pages 41–48, Ann Arbor, MI, June.

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

Sepesy Maučec, Mirjam, Janez Brest, and Zdravko Kačič. 2006. Slovenian to English Machine Translation using Corpora of Different Sizes and Morpho-syntactic Information. In *Proceedings of the 5th Language Technologies Conference*, pages 222–225, Ljubljana, Slovenia, October.

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available Translation Memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*, pages 454–459, Istanbul, Turkey, May.

Stolcke, Andreas. 2002. SRILM – an extensible language modeling toolkit. volume 2, pages 901–904, Denver, CO, September.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*, pages 2214–2218, May.

Toral, Antonio, 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 CroatianEnglish for the Tourism Domain. In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, pages 221–224, Dubrovnik, Croatia, June.

Tyers, Francis M. 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.

Žganec Gros, Jerneja and Stanislav Gruden. 2007. The voiceTRAN machine translation system. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 07)*, pages 1521–1524, Antwerp, Belgium, August. ISCA.