Enlarging scarce in-domain English-Croatian corpus for SMT of MOOCs using Serbian

Maja Popović, Kostadin Cholakov, Valia Kordoni, Nikola Ljubešić*

Humboldt University of Berlin, Germany
name.surname@hu-berlin.de

* Dept. of Knowledge Technologies, Jožef Stefan Institute, Slovenia
nikola.ljubesic@ijs.si

Abstract

Massive Open Online Courses have been growing rapidly in size and impact. Yet the language barrier constitutes a major growth impediment in reaching out all people and educating all citizens. A vast majority of educational material is available only in English, and state-of-the-art machine translation systems still have not been tailored for this peculiar genre. In addition, a mere collection of appropriate in-domain training material is a challenging task. In this work, we investigate statistical machine translation of lecture subtitles from English into Croatian, which is morphologically rich and generally weakly supported, especially for the educational domain. We show that results comparable with publicly available systems trained on much larger data can be achieved if a small in-domain training set is used in combination with additional in-domain corpus originating from the closely related Serbian language.

1 Introduction

Massive Open Online Courses (MOOCs) have been growing rapidly in size and importance, but the language barrier constitutes a major obstacle in reaching out all people and educating all citizens. A vast majority of materials is available only in English, and state-of-the-art machine translation (MT) systems still have not been tailored for this type of texts: the specific type of spoken language used in lectures, ungrammatical and/or incomplete segments in subtitles, slides and assignments, a number of distinct courses i.e. domains such as various natural sciences, computer science, engineering, philosophy, history, music, etc.

Machine translation of this genre into an under-resourced morphologically rich target language represents an additional challenge – in this work, we investigate translation into Croatian. Croatian has recently become the third official South Slavic language in the EU,[1] but it is still rather under-resourced in terms of free/open-source language resources and tools, especially in terms of parallel bilingual corpora. Finding appropriate parallel educational data is even more difficult. Therefore, we based our experiments on a small in-domain parallel corpus containing about 12k parallel segments. We then investigate in what way the translation quality can be improved by an additional in-domain corpus of about 50k segments containing a closely related language, namely Serbian. In addition, we explore the impact of adding a relatively large (200k) out-of-domain news corpus.

Croatian and Serbian are rather close languages, so one option could be to directly use additional English-Serbian data. However, previous work has shown a significant drop in translation quality for a similar cross-language translation scenario (Popović and Ljubešić, 2014). Therefore we also investigate a high-quality Serbian-to-Croatian rule-based MT system for creating additional artificial English-Croatian data.

[1] together with Slovenian and Bulgarian

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 97–105, Osaka, Japan, December 12 2016.

1.1 Related work

In the last decade, several SMT systems have been built for various South Slavic languages and English. Through the transLectures project,[2] transcriptions and translation technologies for automatic generation of multilingual subtitles for online educational data were provided for a set of language pairs. Through this project, one of the South Slavic languages, namely Slovenian, became an optional language pair in the 2013 evaluation campaign of IWSLT (International Workshop on Spoken Language Translation) (Cettolo et al., 2013). The SUMAT project[3] included translation between Serbian and Slovenian subtitles (Etchegoyhen et al., 2014), however translation from English was not explored and no educational subtitles were used.

For Croatian-English language pair, first results are reported in (Ljubešić et al., 2010) on a small weather forecast corpus. Translation between Croatian and English has become one of the focuses of the AbuMatran project:[4] an SMT system for the tourist domain is presented in (Toral et al., 2014), the use of morpho-syntactic information by means of factored models is investigated in (Sánchez-Cartagena et al., 2016) and several scenarios with different models and data-sets are explored in (Toral et al., 2016). Furthermore, SMT systems for the news domain for Croatian and Serbian are described in (Popović and Ljubešić, 2014).

To the best of our knowledge, no systematic investigation on English-to-Croatian educational data has been carried out yet.

2 Challenges for machine translation of MOOCs

As already mentioned, machine translation of MOOCs induces several challenging tasks. A first step for building a statistical machine translation (SMT) is collection of parallel data. For educational genre, already this step is a challenge due to the following reasons:

- crawling: the structure of the web resource containing desired material is complex and does not allow for large-scale automatic crawling;
- data extraction and alignment: a large portion of materials is in pdf format, which can lead to misalignments during conversion into plain text;
- the size of extracted data: the available data is often very small, so that machine translation with scarce resources has to take place;
- target languages: majority of materials is available in English, meaning that machine translation into morphologically rich languages has to take place;
- representativeness: majority of available materials are lecture subtitles; slides, notes and assignments are unfortunately rarely translated.
- copyright issues are often not clear and are difficult to define.

Once the parallel data is extracted from some source, segmentation can be a challenging task itself. The platforms used for translation are primarily designed for subtitles so that the translators are encouraged to use short segments which often represent incomplete and/or ungrammatical sentences. Another peculiarity is the fact that the lecturers often do not finish a sentence properly or change the subject in the middle of a utterance.

3 Challenges for English-Croatian machine translation and getting help from Serbian

Croatian, as a Slavic language, has a very rich inflectional morphology for all word classes. There are six distinct cases, three genders and a number of verb inflections since person and many tenses are expressed

[2] https://www.translectures.eu/
[3] http://www.sumat-project.eu/
[4] http://www.abumatran.eu/

by the suffix. In addition, negation of three important verbs is formed by adding the negative particle to the verb as a prefix. As for syntax, the language has a quite free word order, and there are no articles, neither indefinite nor definite. In addition, multiple negation is always used.

All these morpho-syntactic peculiarities are even more difficult to generate correctly if the available resources are scarce, especially for spoken language style used in lectures as well as for ungrammatical and/or unfinished sentences.

Differences between Croatian and Serbian Both languages belong to the South-Western Slavic branch. Although they exibit a large overlap in vocabulary and a strong morpho-syntactic 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.

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: constructions involving modal verbs, especially those with the verb "trebati" (to need, should), future tense, conditional.

4 Research questions

Taking into account the facts described in previous sections, i.e. peculiarities of educational genre, difficulties regarding characteristics of the Croatian language and scarceness of available resources, as well as similarities and differences between Croatian and Serbian, our main questions are:

- how does the translation performance for a small in-domain training data compare with the performance for a larger out-of-domain data?
- is it possible to increase the performance by adding Serbian in-domain data and what is the optimal way?

Our work is to certain extent related to the experiments described in (Popović and Ljubešić, 2014). They explored adaptation of the in-domain news test data to the desired language by applying a set of simple rules or a Serbian-Croatian SMT system, whereas the training data for the English-Serbian/Croatian system were fixed. We investigate different combinations of training data for the challenging genre i.e. educational material in order to build a ready SMT so that the test data once translated do not require any further intervention. In addition, we use a recently developed high quality rule-based Serbian-to-Croatian system (Klubička et al., 2016) which performs better than SMT systems used in (Popović and Ljubešić, 2014).

5 Experimental set-up

Parallel texts

The data used in our experiments are collaboratively translated subtitles from Coursera[5] and contain several types of courses/domains: biology, computer science, philosophy, nutrition, music, etc. The translations are produced by course participants who usually translate into their native languages. Translations are done via a collaborative platform which is usually used for translation of movie subtitles, thus not designed with large-scale crawling in mind. In order to crawl the relevant data, we first had to construct manually a list of the Coursera courses available there. Once the list of translated Coursera courses was constructed, Python scripts were used to download the original English data and the corresponding translations. However, this process was not fully automatic because there were some issues with the format of the URLs of some of the courses as well as the data format of the translations.

[5]https://www.coursera.org/

The parallel data collected is of a relatively good quality. The texts are mostly properly aligned, however the sentence segmentation is not optimal. As mentioned in Section 2, the extracted parallel segments often contain incomplete sentences or parts of two different sentences. Of course, one can think of automated correction of segmentation. However, for bilingually aligned texts this represents a peculiar task for several reasons:

- there are no apparent punctuation rules which are consistent in both languages: some sentences end with "." i one language but with "," or ";" or conjunction or even nothing in the other;
- some consecutive English source segments are only partially translated (Table 1) – if these segments were merged in both languages, a proper English sentence aligned with an incorrect and/or ungrammatical translation would be generated.

English	Croatian
Five years ago	Pre pet godina
I was told specifically	(no translation)
this is his name	da mu je to ime.

Table 1: Example of English successive segments and their Croatian translations: the middle of the sentence is not translated at all.

For these reasons, and also taking into account the fact that the test set is in the same format, no resegmentation attempts were performed and the texts are used directly in the format they were extracted. Nevertheless, since the data are not completely clean, certain filtering steps had to be performed. First of all, there was a large number of short redundant segments such as "Mhm", "Hello", "Welcome back", etc. These segments were separated from the rest according to the sentence length and only a unique occurrences were kept. The rest of the corpus was then cleaned from incorrect translations on the base of sentence length: if the proportion of source and target sentence length was too high or too small, the segment was removed. As a final step, the two cleaned parts of the corpus were merged. The same procedure was carried out for both for English-Croatian as well as for English-Serbian data sets. For English-Croatian, about 12k parallel segments were extracted, and for English-Serbian about 50k. An interesting observation is that although Croatian is generally better supported in terms of publicly available parallel data,[6] Serbian is currently better supported for educational parallel texts.

As for the out-of-domain corpus, we used the SETimes news corpus (Tyers and Alperen, 2010) since it is relatively large (200k parallel sentences) and clean.

Moses set-ups

We trained the statistical phrase-based systems 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 SRILM (Stolcke, 2002).

The investigated bilingual training set-ups are:

1. en-hr SEtimes (relatively large clean out-of-domain corpus)
2. en-hr Coursera (small in-domain corpus)
3. en-hr Coursera (small in-domain corpus) + en-sr Coursera (larger in-domain corpus)
4. en-hr Coursera + en-hr' Coursera
5. en-hr SEtimes + en-hr Coursera + en-hr' Coursera

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

		sentences	running words		voc		oov (%) (dev/test)	
			en	hr	en	hr	en	hr
Training	1) setimes	206k	4.9M	4.6M	68k	137k	2.7/2.4	10.9/7.4
	2) coursera	12k	148k	118k	8k	17k	5.5/5.5	8.2/8.8
	3) 2+coursera_en-sr	62k	782k	659k	21k	54k	1.5/1.2	5.3/5.7
	4) 2+coursera_en-hr'	62k	782k	696k	21k	52k	1.5/1.2	4.9/5.2
	5) 1+4	268k	5.7M	5.3M	76k	162k	0.8/0.6	2.9/2.9
Dev	coursera	2935	28k	23k	3.8k	6.3k		
Test	coursera	2091	25k	20k	3.4k	5.5k		

Table 2: Data statistics.

where hr' denotes Serbian part of the corpus translated by a rule-based machine translation system into Croatian. For each set-up, the language model was trained on the target part of the used bilingual corpus. For set-ups including combined parallel corpora (3, 4 and 5), the corpora were merged by simple concatenation and the interpolated language model was used. Data statistics for all set-ups can be seen in Table 2.

Serbian-to-Croatian RBMT system

The MT system (Klubička et al., 2016) used for creating additional artificial Croatian data from Serbian is a bidirectional rule-based system based on the open-source Apertium platform (Forcada et al., 2011). Considering the fact that differences between Croatian and Serbian occur mostly at the lexical and orthography, using a rule-based system makes the most sense. The system tested on newspaper texts achieves 83.0% BLEU for translation into Croatian, whereas the BLEU score is 72.7% if the Serbian source is directly compared to the Croatian reference translation.

Evaluation

For all set-ups, BLEU scores (Papineni et al., 2002) and character n-gram F-scores i.e. CHRF3 scores (Popović, 2015) are reported. In addition, five Hjerson error classes (Popović, 2011) are reported in order to get a better insight into differences between the systems: inflectional errors, ordering errors, missing words, additions and lexical errors.

6 Results

6.1 Automatic evaluation scores

Table 3 presents the obtained automatic scores for all Moses training set-ups described in Section 5 together with the scores for translations generated[7] by two publicly available SMT systems for English-to-Croatian: Asistent[8] (Arčan et al., 2016) and Google translate[9].

It can be seen that the most promising set-up according to automatic evaluation metrics is the set-up 5, i.e merging both domains and adding artificial in-domain English-Croatian parallel text where the target Croatian part is generated from Serbian by the rule-based MT system. This set-up even outperforms the Asistent system which is trained on much larger parallel texts, albeit none of them from educational domain.

Furthermore, it can be seen that both SETimes and original Coursera set produce the same percentage of lexical errors – the first one due to the domain discrepance and the other due to data sparsity. Adding in-domain Serbian data reduces the number of lexical errors, which is further reduced by translating Serbian into Croatian. Merging of two data-sets reduces lexical errors even more, however their number is still larger than for Asistent and Google systems.

[7] in June 2016

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

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

	overall scores		Hjerson error rates					
system	BLEU	CHRF3	infl	order	miss	add	lex	Σer
1) setimes	8.1	38.5	10.6	5.0	6.4	10.5	40.8	73.2
2) coursera	12.7	38.9	7.5	4.2	4.0	14.6	40.8	71.1
3) 2+coursera-sr	13.2	41.1	8.8	4.7	5.3	11.8	38.4	69.2
4) 2+coursera-hr'	14.1	42.6	9.4	4.8	5.3	11.8	37.0	68.4
5) 1+4	**15.5**	**44.9**	10.2	5.0	6.5	9.9	35.5	**67.1**
asistent	14.7	43.5	9.9	5.2	8.1	9.4	34.7	67.4
google	17.1	49.4	8.2	4.5	4.4	13.8	30.1	61.0

Table 3: Automatic evaluation scores (%) for each of the systems: BLEU score, CHRF3 score and five Hjerson error rates: inflectional, ordering, omission, addition and lexical error rate together with their sum.

Ordering errors and omissions are lower for the set-ups without SEtimes, most probably due to different sentence (i.e. segment) structure in two genres/domains.

Morphological errors are also lower without SEtimes, however they are high in all set-ups which should be generally addressed in future work by using morpho-syntactic analysers and/or generators.

Apart from this, it can be observed that the main advantage of the Google system is the low number of lexical errors which is probably achieved by using very large training corpora.

6.2 Translation examples

In order to illustrate advantages and disadvantages of different SMT systems, Table 4 shows six English source segments and their translations by each of the systems. Erroneous parts of the obtained translations are annotated by parentheses: {} stands for lexical errors, additions, omissions and inflections (where only part of the word is in parenthesis), // stands for ordering errors and <> for stylistic variants.

segment 1: A completely correct sentence is produced only by the set-up 5, as well as by the publicly available systems. The other systems generate ungrammatical sentences, en-hr Coursera alone generates stylistically questionable translation.

segment 2: None of the systems produces a perfect translation – however, the most accurate translation containing only two minor morphological errors is produced by set-up 5, i.e. combination of all Coursera data and SETimes.

segment 3: Spoken lecture language issues: SEtimes produces the worst translation, followed by Google and then Asistent; all set-ups with Coursera data produce correct translations.

segment 4: The translation of the incomplete segment is difficult for all systems. Both SETimes and Croatian Coursera alone generate very bad translations – the first one because of domain discrepancy, the second one because of data sparsity; the other set-ups generate ungrammatical segments where the meaning still can be captured. Asistent produces the best translation containing only one inflectional error which does not change the meaning.

segment 5: The best translation (without any error) of another incomplete sentence is generated by set-up 4, i.e. Coursera with additional artificial data; the worst translation is generated by SEtimes, which also introduces morphological errors when combined with Coursera in set-up 5. This example illustrates that in-domain data are important not only for vocabulary and lexical errors but also for morpho-syntactic properties.

segment 6: Spoken language and incomplete sentence: SETimes, Google and Asistent produce a number of errors; using Serbian instead of Croatian induces some errors mainly due to differences in verb structures; the best option is the use of Croatian Coursera with or without additional data.

1)	Is this a problem?
setimes	Je {} to problem{a}?
coursera	Da li <bi> to <bio> problem?
coursera+sr	Je {} <bi> to <bio> problem?
coursera+hr'	/Li/ <bi> to <bio> problem?
all	Je li to problem?
asistent	Je li to problem?
google	Je li to problem?
2)	Then the next thing we need, is energy.
setimes	{Tada} sljedeća stvar {} nam treba, je energija.
coursera	{Onda} sljedeće {} trebamo je energije.
coursera+sr	Sljedeća stvar koju {moramo}, je energije.
coursera+hr'	Sljedeća stvar koju {moramo} je energije.
all	Sljedeća stvar koj{u} nam treba, je energij{e}.
asistent	{onda} sljedeća stvar koju trebamo, je energija.
google	{onda je} sljedeća stvar koju trebamo, je energija.
3)	Now on an Android device, of course yeah, there is no mouse.
setimes	{Sada} {o} Android <naprava>, naravno {yeah}, nema miša.
coursera	Na Android uredjaju, naravno, nema miša.
coursera+sr	Na Android uredjaju, naravno, nema miša.
coursera+hr'	Na Android uredjaju, naravno, nema miša.
all	Na Android uredjaju, naravno, nema miša.
asistent	{Sada} Android uredjaj, naravno, nema miša.
google	{Sada} na Android uredjaju, naravno da{,} nema miš{}.
4)	but the number of insects flying in the
setimes	ali broj {insects} {} lete{nja} u
coursera	ali broj insek{ti} {} {flying} u
coursera+sr	ali broj insek{ti} {} lete{ći} u
coursera+hr'	ali broj insek{ti} {} lete{ći} u
all	ali broj insek{ti} {} lete{ći} u
asistent	ali broj insek{tima} koji lete u
google	ali broj insekata {} let{i} u
5)	that he learned French and English later in life.
setimes	{kako} nauči{la} francusk{e} i englesk{e} kasnije u životu.
coursera	da je naučio francu{zi} i {jezikom} {later} u {life}.
coursera+sr	da je naučio francuski i engleski kasnije u životu.
coursera+hr'	da je naučio francuski i engleski kasnije u životu.
all	da je naučio francusk{e} i englesk{e} kasnije u životu.
asistent	{to} je naučio francuski i engleski kasnije u životu.
google	{koji} je kasnije u životu naučio francuski i engleski.
6)	Now you can see up here that I need to select Android mode, so these are the
setimes	{Sada} možete vidjeti {iz} ovdje {kako} trebam izabrati android način, tako {} to /su/
coursera	Ovdje gore možete vidjeti da trebam odabrati android način rada. dakle, ovo
coursera+sr	Ovdje gore možete videti {koji} <ja> treba da izabere{te} android način rada. znači ovo su
coursera+hr'	Ovdje gore možete vidjeti da trebam odabrati android način rada. dakle, ovo
all	Ovdje gore možete vidjeti da trebam odabrati android način rada. dakle, ovo
asistent	{Sada} možeš vidjeti {što} ja moram odabrati android {}, {} ovo su
google	{Sada} možete vidjeti ovdje da moram odabrati android mod{u}, tako da su to

Table 4: Examples of six English source sentences and their translations by different SMT system set-ups; erroneous parts are annotated by {} (mistranslations, additions, omissions, inflections), // (order) and <> (style).

It can be noted that the additional Serbian data does not help only in the first example – only when a larger out-of-domain data is added, a correct translation is obtained. For segment 2), the baseline English-Croatian corpus already yielded a correct translation and there is no change when any of additional corpora is used. In example 3) both the direct use of Serbian and translating into Croatian help to some extent but some errors are still present. For segments 4) and 5) both untranslated and translated Serbian texts result in the same correct translation. For the example 6) using the translated additional data significantly improves the performance in comparison with the "raw" Serbian data.

7 Summary and outlook

This work has shown that a small amount of in-domain training data is very important for the English-to-Croatian statistical machine translation of the specific genre of Massive Open Online Courses, especially for capturing appropriate morpho-syntactic structure. Adding in-domain data containing the closely related Serbian language improves the performance, especially when the Serbian part is translated into Croatian thus producing an artificial English-Croatian in-domain corpus. The improvements consist mainly from reducing the number of lexical errors. Further improvements have been achieved by adding a relatively large out-of-domain news corpus reaching performance comparable with systems trained on much larger (out-of-domain) parallel texts. Adding this corpus reduces the number of additions and lexical errors, nevertheless it introduces more morphological and ordering errors due to the different nature and structure of the segments.

Future work should include investigating better ways of combining and extracting relevant information from original (in-domain) and additional (out-of-domain and/or "out-of-language") data. In addition, the use of morpho-syntactic information should be explored, especially since this also represents a challenging task for the peculiar genre such as educational material.

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. 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).

References

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

Mauro Cettolo, 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.

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.

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.

Nikola Ljubešić, Petra Bago, and Damir Boras. 2010. Statistical machine translation of Croatian weather forecast: How much data do we need? In Vesna Lužar-Stiffler, 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.

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.

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 M. Sánchez-Cartagena, Nikola Ljubešić, and Filip Klubička. 2016. Dealing with data sparseness in SMT with factored models and morphological expansion: a Case Study on 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.

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

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 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.

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.