PE2 rr corpus: manual error annotation of automatically pre-annotated MT post-edits

Title	PE2 rr corpus: manual error annotation of automatically pre-annotated MT post-edits
Author(s)	Popovic, Maja;Arcan, Mihael
Publication Date	2016-05-23
Publisher	European Language Resources Association

PE^{2}rr Corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits

Maja Popović
Humboldt University of Berlin
Berlin, Germany
maja.popovic@hu-berlin.de

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

Abstract

We present a freely available corpus containing source language texts from different domains along with their automatically generated translations into several distinct morphologically rich languages, their post-edited versions, and error annotations of the performed post-edit operations. We believe that the corpus will be useful for many different applications. The main advantage of the approach used for creation of the corpus is the fusion of post-editing and error classification tasks, which have usually been seen as two independent tasks, although naturally they are not. We also show benefits of coupling automatic and manual error classification which facilitates the complex manual error annotation task as well as the development of automatic error classification tools. In addition, the approach facilitates annotation of language pair related issues.

Keywords: machine translation, post-editing, error annotation

1. Introduction

The amount of post-edited translation outputs as well as the amount of data containing annotated machine translation errors is growing in the recent years, with more and more applications and growing need for such data. The process of improving a machine-generated translation is a natural task – post-edited translations are a by-product of the professional translation workflow, whereas (explicit) translation error classification, which is a very difficult task, is not. On the other hand, the two tasks are actually highly related – post-editing can be viewed as implicit error annotation, since each edit operation is actually a correction of a translation error. Despite of this fact, these two tasks have almost always been seen and performed completely separated and independent. Therefore, data containing error annotations of actual post-edit operations are scarce. The corpus PE^{2}rr aims to improve this situation, containing post-edited translation outputs where each post-edit operation is assigned to a particular error class.

In addition, another discrepance in the error classification experiments and corpora can be observed: manual error classification and automatic error classification have always been carried out completely separately from each other. Whereas the results of manual error annotation present a valuable resource for development of automatic error classification tools, this independent approach makes detailed analysis of drawbacks of automatic tools difficult. The PE^{2}rr corpus has been created by merging these two tasks together: automatic error classification has been performed as a first step, and the manual error classification has been performed as a correction of these automatically pre-annotated errors. A very important advantage of this approach is also reducing the effort of the complex and time-consuming manual error classification process, since correcting pre-annotated errors is easier and faster than looking for the errors and defining them from scratch. We believe that the approach can also improve the inter-annotator (and intra-annotator) agreement, however this aspect will be investigated in future work.

The benefits for the community can be manifold:

- facilitating improvement and development of automatic error classification and error prediction tools;
- further usage as additional post-edited and/or error annotated data for quality estimation and error prediction, automatic-post-editing, etc.;
- investigating correlations between automatic evaluation metrics and certain errors;
- indicating most prominent problems of the state-of-the-art SMT systems for the described target languages as well as for a number of related ones.

We also believe that our positive experience of joining post-editing, automatic and manual error classification will be further used for more language pairs and by professional translators, and that the community will derive valuable knowledge about the annotation process and translation errors from the corpus we presented.

1.1. Related work

Publicly available post-edited data have been used for a while for quality estimation and error prediction WMT tasks[1], also for automatic post-editing task (Bojar et al., 2015). Parts of the corpora are also error annotated, though only with basic edit distance operations (substitution, deletion, insertion and shift) or binary tags ("ok" for correct words and "bad" for erroneous ones). Another corpus containing same type of edit annotation is the TRACE corpus (Wisniewski et al., 2013) which consists of French-English and English-French post-edited translation outputs annotated with basic edit distance error types.

Detailed manual error analysis has been often used in the last decade to determine the most prominent errors for particular task/translation system (e.g. (Vilar et al., 2006) at

[1] http://www.statmt.org/wmt15/quality-estimation-task.html

the beginning, (Lommel et al., 2014) recently), as well as to investigate impacts of different error classes to various aspects of translation quality ((Kirchhoff et al., 2012; Federico et al., 2014). Nevertheless, none of these error analyses have been carried out on post-edits, only on raw translation outputs.

Manual error analysis of Finish post-edits has been carried out in (Koponen, 2013) in order to investigate discrepances between the estimated cognitive effort and actual technical effort, however, to the best of our knowledge, the corpus is not publicly available.

The publicly available TARAXÜ corpus (Avramidis et al., 2014) contains both post-edited as well as error annotated translation outputs, whereby the two tasks were carried out completely separately, and even not on the same set of translation outputs.

The Terra corpus (Fishel et al., 2012) is a publicly available collection of manually error annotated corpora[2] which has been used for assessment of the automatic classification tools Addicter (Zeman et al., 2011) and Hjerson (Popović, 2011). The corpora were annotated by different research teams independently and the annotation strategies were different: from a free annotation using only the source text and the translation output without taking into account any reference translation, to a flexible reference-based annotation where a reference translation has also been taken into account. Although this corpus has been very valuable for automatisation of error classification, all data sets pose difficulties for detailed and precise assessment of automatic tools and their further development, because (i) standard reference translations differ much more from translation outputs than post-edits, so that automatic error annotation tools tag a large number of actually correct words and (ii) manual and automatic error classification were performed completely independently.

2. Data sets

The corpus PE^{2}rr contains Serbian and Slovenian SMT generated translations from English and German source texts, as well as a part of the TARAXÜ corpus, namely German, Spanish and English SMT generated translations from English, Spanish and German source texts. The Serbian and Slovenian languages, as Slavic languages, have quite free word order and are highly inflected. The derivational morphology is also rich, multiple negation is used, and there are no articles, only determiners. Similarly, German is also a morphologically rich language and a very challenging language for machine translation. Spanish is generally less inflective than the Slavic languages and German, however the number of possible verb inflections is rather high. The following domains/genres were used: news texts from the enhanced version[3] of the SETimes (Tyers and Alperen, 2010) corpus for English→Serbian, EuroParl (Koehn, 2005) for English/German→Slovenian and OpenSubtitles[4] for all language pairs. All the corpora are downloaded from the OPUS web site[5] (Tiedemann, 2012). It should be noted that the OpenSubtitles corpus contains transcriptions and translations of spoken language thus being slightly peculiar for machine translation. From the TARAXÜ corpus, the post-edited part containing WMT News texts is used.

Table 1 gives an overview of the total amount of sentences and words used for generation of the corpus PE^{2}rr for each domain and translation direction.

All translations have been generated using phrase-based Moses (Koehn et al., 2007), where the word alignments were built with GIZA++ (Och and Ney, 2003). The 5-gram language model was built with the SRILM toolkit (Stolcke, 2002). For each translation output, a system was trained on the corresponding in-domain parallel data.

3. Post-editing and error annotation

For the Serbian and Slovenian MT outputs, both tasks, i.e. post-editing and error annotation, were performed by MT researchers with some experience with human translation. The annotators are native speakers of the target languages and fluent in both source languages. For the TARAXÜ data, post-editing was performed by professional translators and error annotation was performed by MT researchers fluent both in the corresponding source and target languages.

The evaluation has been carried out in three steps: first, each machine generated translation has been post-edited. Then, automatic error classification has been applied to assign an error category to each post-edit operation. Finally, manual inspection and correction of these error labels has been carried out.

In addition to standard error classification, first steps towards the annotation of the most prominent language(-pair) related issues (Popović and Arčan, 2015) has been carried out.

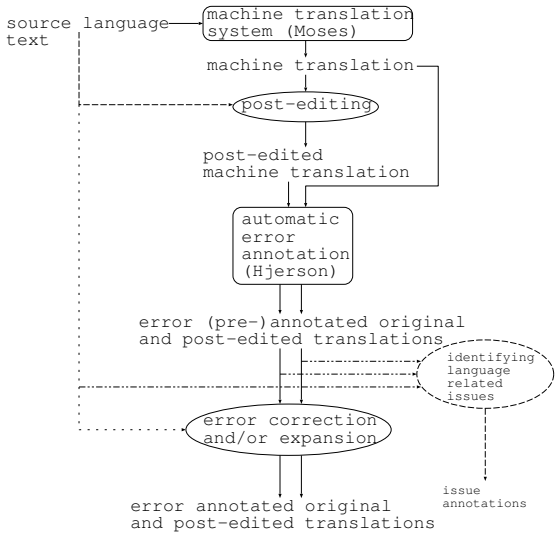

Figure 1: Procedure of generating the PE^{2}rr corpus.

The general procedure for generation of the PE^{2}rr corpus is shown in Figure 1. Rectangle processes were carried out automatically and ellipse processes manually.

[2] http://terra.cl.uzh.ch/terra-corpus-collection.html

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

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

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

translation direction	domain	seg	source words	target language texts words	target language texts edits (%)
English→Serbian	SEtimes	300	7311	6848	2080 (30.4)
	OpenSub	440	4352	3589	1157 (32.2)
German→Serbian	OpenSub	440	4182	3490	1171 (33.6)
English→Slovenian	EuroParl	100	3008	2664	397 (14.9)
	OpenSub	440	4352	3702	1201 (32.4)
German→Slovenian	EuroParl	100	2553	2177	495 (22.7)
	OpenSub	440	4182	3619	1125 (31.1)
English→German	News	255	5582	5588	2575 (46.1)
Spanish→German	News	101	2411	2255	902 (40.0)
German→Spanish	News	40	961	1040	485 (46.7)
German→English	News	240	5044	5198	1554 (29.9)
total		2896	43938	40170	13142 (32.7)

Table 1: Number of source language segments and running words together with number of generated target language segments, running words and post-edited words for each domain and each translation direction.

	technical effort (edit distance)			
text	none 0	low 0-25%	medium 25-50%	large >50%
en-sr SEtimes	7.0	28.0	46.7	17.3
en-sr OpenSub	27.0	16.8	30.7	25.4
de-sr OpenSub	28.0	14.3	28.6	29.1
en-sl EuroParl	18.0	60.0	18.0	4.0
en-sl OpenSub	29.1	16.2	27.7	27.0
de-sl EuroParl	11.0	41.0	41.0	7.0
de-sl OpenSub	25.0	18.4	31.6	25.0
en-de News	2.7	9.8	35.7	51.8
es-de News	4.0	15.8	52.5	27.7
de-es News	7.5	15.0	20.0	57.5
de-en News	5.0	31.7	45.0	18.3
total	19.2	20.6	34.0	26.2

Table 2: Distribution of target language sentences (%) according to performed technical effort (edit distance).

3.1. Post-editing

Post-editing has been performed to create a fully fluent and adequate translation which is of the same quality level as a standard human translation. Only a minimal post-editing necessary to achieve an acceptable translation quality (adequate and fluent translation) has been performed, without taking into account potential editors' preferences concerning style, lexical choice, grammatical structure, etc. Table 1 shows number of segments and running words in all source texts and in all generated machine translation outputs together with the number of post-edited words (which is about 30% on average).

Apart from that, a division of performed technical effort into four categories is presented in Table 2. It can be seen that overall, about one third of the segments required medium technical effort, 20% segments required only small intervention, 20% were already acceptable, and about 25% required significant modifications. Edit distances for each segment in the form of WER are a part of the corpus, too.

3.2. Error annotation

The task of error annotation consisted of assigning an error class to each post-edit operation and was performed in two stages. The first stage consisted of automatic pre-annotation by Hjerson (Popović, 2011), a tool for automatic error classification which enables categorisation into five error classes: addition, inflectional error (verb tense/person/mood, case, gender, number), lexical error (mistranslation), omission, and word order error. For Slovenian translations, lemmas were generated by Obelix (Grčar et al., 2012). For Serbian translations, "poor man's" version of Hjerson (Popović et al., 2015), based on first four letters of words instead of lemmas, was used. For the TARAXÜ target languages, the TreeTagger[6] was used. The second stage consisted of correcting or expanding these automatically generated error classes. Three additional error classes which were observed as recurrent and frequent were introduced:

- contraction
  separated, incorrect or missing (parts of) compounds/contractions:
  Serbian and Slovenian negation particle+verb contractions, German preposition+article contractions, Slovenian preposition+pronoun contractions, Spanish article+preposition and verb+pronoun contractions, as well as different types of compounds in all languages.

- derivation
  incorrect POS, verb aspect, verb prefix, passive/past-participle confusion, possessive adjectives and pronouns

- untranslated
  out-of-vocabulary words in the source language which remained untranslated in the output

Apart from that, some errors can be assigned to more than one error class – a number of contractions, derivations, inflections, mistranslations and unknown words can be placed

[6] http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/

at an incorrect position in a sentence, i.e. they are also reordering errors. This phenomenon is not possible to detect by the state-of-the-art automatic error classification tools so that these multiple tags were assigned during the manual evaluation.

3.2.1. Distribution of manual error classes

Distribution of error classes in the corpus PE^{2}rr are presented in Table 3 in the form of raw counts and error rates for each of the error categories. The error rates are obtained by dividing raw counts with total number of running words and therefore are suitable for comparison of different translation outputs.

3.2.2. Annotation of language related issues

Different language combinations exhibit different error distributions in the translation output which often relates to linguistic characteristics of involved languages as well as to divergences between them. Those relations have been investigated in (Popović and Arčan, 2015) and 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 domain and text type, mostly differing for written and spoken language.

Manual inspection of source sentences and their corresponding machine translations annotated by Hjerson using the independent reference translations together with their corresponding source sentences has been carried out in these experiments, and the phenomena were only counted, not annotated.

In this work, we carried out a preliminary experiment of annotating the part of the PE^{2}rr corpus with these issues. The labels were assigned on the segment level and the distributions of the ten most frequent issues for each annotated text are presented in Table 4.

Finding the best definition of issues and the best method for annotating is still a part of the ongoing research, but detecting and defining an issue is certainly much easier when inspecting errors in the form of edit operations instead of errors with respect to the independent references.

3.3. Comparison between manual and automatic error annotation results

The comparison between Hjerson counts and manual classification counts has been carried out in the same way as in (Popović and Burchardt, 2011), namely:

- comparison of error distributions within one translation output as well as across different translation outputs in the form of Spearman correlation coefficients;
- precision, recall and confusions between different error classes.

de-en News	Hjerson error class					
	add	infl	lex	miss	ord	x
prec. PE^{2}rr	58.1	84.2	67.4	61.5	88.4	99.2
prec. Terra	29.0	37.5	29.6	48.4	15.3	96.0
recall PE^{2}rr	56.0	91.8	84.6	62.2	95.5	98.9
recall Terra	16.7	92.3	85.8	45.8	51.4	55.3

Table 5: Precision and recall for Hjerson error categories for the German→English News part of the PE^{2}rr corpus compared with the similar text from the Terra corpus. Label "x" stands for correct words (not edited/no error).

The correlations were already very high in the experiments reported in (Popović and Burchardt, 2011), but using the PE^{2}rr corpus resulted in a significant increase in recalls and especially precisions for all error classes. Table 5 presents precision and recall for Hjerson classes with regard to manual classes. The results are reported for the German→English News part of the PE^{2}rr corpus and compared with the same text type[7] from the Terra corpus analysed in (Popović and Burchardt, 2011) where independent reference translations were used for automatic error classification and the results are afterwards compared with results of a completely independent manual error analysis. This means that the PE^{2}rr corpus, where the reference translations are post-edited translation outputs and a manual error annotation has been carried out on pre-annotated texts, can give much better and more reliable insights into particular flaws of an automatic error classification tool.

A detailed analysis of Hjerson performance, being an important direction for future experiments, is however out of the scope of this work. Nevertheless, it should be already noted that Hjerson definitely exhibits a significant number of confusions between lexical errors, omissions and additions – addressing this problem should be one of the first steps for its improvement.

4. Summary and Outlook

We have presented a freely available corpus[8] containing automatically generated translations into several highly inflective languages, their post-edited versions, and error annotations of the performed post-edit operations. We believe that the corpus will be useful for many distinct applications. The main advantage of the approach used for creation of the corpus is the fusion of post-editing and error classification tasks, which have usually been seen as two independent tasks, although naturally they are not. In addition, we show benefits of coupling automatic and manual error classification which facilitates the complex manual error annotation task as well as the development of automatic error classification tools.

Future work on the presented corpus includes adding more language pairs and domains to the collection, further annotation and investigation of the language related obstacles, and further in-depth analysis and improvement of the state-of-the-art methods for automatic translation error analysis.

[7] Language pair, machine translation system type, translation direction and domain are the same, the exact sentences are not.

[8] http://nlp.insight-centre.org/research/resources/pe2rr/

(a) Raw error counts

text		raw counts for each error class													
		add	contr	+ord	der	+ord	infl	+ord	lex	+ord	miss	ord	+else	untr	+ord
en-sr	SEtimes	237	11	+3	107	+2	751	+59	382	+18	264	222	+83	23	+1
	OpenSub	114	9	+0	44	+3	270	+11	335	+11	253	97	+25	10	+0
de-sr	OpenSub	111	9	+2	27	+3	203	+28	289	+18	275	142	+61	54	+10
en-sl	EuroParl	55	3	+0	16	+0	133	+5	83	+2	60	38	+7	2	+0
	OpenSub	100	12	+5	53	+3	291	+17	305	+10	259	135	+35	11	+0
de-sl	EuroParl	62	2	+0	7	+0	121	+9	103	+9	128	32	+22	19	+3
	OpenSub	97	15	+1	21	+1	238	+33	273	+29	204	153	+73	51	+9
en-de	News	280	263	+20	59	+19	374	+53	544	+117	332	448	+222	53	+13
es-de	News	60	66	+2	11	+4	132	+13	152	+22	224	205	+47	9	+6
de-es	News	44	13	+0	4	+1	44	+8	111	+23	113	91	+36	29	+4
de-en	News	141	27	+0	16	+3	84	+26	371	+57	339	401	+108	67	+22
total		1301	430	+33	365	+39	2641	+262	2948	+316	2451	1964	+718	328	+68

(b) Error rates

text		error rates [%] for each error class													
		add	contr	+ord	der	+ord	infl	+ord	lex	+ord	miss	ord	+else	untr	+ord
en-sr	SEtimes	3.46	0.16	+.04	1.56	+.03	11.0	+.86	5.58	+0.26	3.85	3.24	+1.20	0.34	+.01
	OpenSub	3.19	0.25	+0	1.23	+.08	7.54	+.31	9.37	+0.31	6.80	2.71	+0.70	0.28	+0
de-sr	OpenSub	3.19	0.26	+.06	0.78	+.09	5.83	+.80	8.30	+0.52	7.53	4.08	+1.76	1.55	+.29
en-sl	EuroParl	2.06	0.11	+0	0.60	+0	4.99	+.19	3.11	+0.08	2.26	1.43	+0.27	0.08	+0
	OpenSub	2.70	0.32	+.13	1.43	+.08	7.87	+.46	8.24	+0.27	6.73	3.65	+0.94	0.30	+0
de-sl	EuroParl	2.85	0.09	+0	0.32	+0	5.56	+.41	4.73	+0.41	5.67	1.47	+0.96	0.87	+.14
	OpenSub	2.69	0.41	+.03	0.58	+.03	6.58	+.91	7.54	+0.80	5.46	4.23	+2.02	1.41	+.25
en-de	News	5.01	4.71	+.36	1.06	+.34	6.69	+.94	9.73	+2.09	6.12	8.01	+3.96	0.95	+.23
es-de	News	2.66	2.93	+.09	0.49	+.18	5.85	+.58	6.74	+0.98	9.37	9.09	+2.10	0.30	+.27
de-es	News	4.23	1.25	+0	0.38	+.10	4.23	+.77	10.7	+2.21	10.1	8.75	+3.46	2.79	+.38
de-en	News	2.71	0.52	+0	0.31	+.06	1.62	+.50	7.14	+1.10	6.18	7.71	+2.08	1.29	+.42
total		3.24	1.07	+.08	0.91	+.10	6.57	+.65	7.34	+0.78	6.00	4.89	+1.78	0.82	+.17

Table 3: Distribution of error classes: number and percentage of words belonging to each of the eight error classes; counts for multiple error classes are added to each of the corresponding categories by "+n". Percentage is calculated with respect to total number of running words.

SEtimes & OpenSubtitles				EuroParl & OpenSubtitles				News	
en-sr		de-sr		en-sl		de-sl		en-de	
n-infl	62.2	oov	15.5	n-infl	27.3	n-infl	28.6	det-infl	55.0
a-infl	34.0	phrase-struct	12.0	v-infl	18.7	oov	11.0	comp	40.0
v-infl	18.8	n-infl	11.5	a-infl	15.3	v-infl	10.7	phrase-struct	31.0
n-coll	14.6	v-infl	8.0	ppast-infl	8.7	a-infl	9.3	n-infl	23.0
prep	10.2	question	6.5	v-reord-local	7.3	ppast-moss	9.0	a-infl	21.0
ppast-infl	9.2	ppast-miss	6.5	prep	5.7	phrase-struct	7.7	v-infl	20.0
literal	8.2	v-aux-miss	5.5	literal	4.8	v-fin-miss	7.3	prep	20.0
v-aux-miss	7.8	prep	5.0	det-infl	4.7	prep	5.3	v-reord	19.0
pron-ext	6.0	literal	5.0	v-aux-miss	4.0	comp-oov	5.3	det-miss	17.0
v-reord-local	5.4	v-aux-ext	4.0	v-aux-ext	4.0	v-aux-ext	5	oov	14.0

Table 4: Distribution of segment level language related issues (normalised with the total number of segments) for five translation outputs.

5. Acknowledgements

This publication has emanated 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).

6. Bibliographical References

Avramidis, E., Burchardt, A., Hunsicker, S., Popović, M., Tscherwinka, C., Torres, D. V., and Uszkoreit, H. (2014). The taraXÜ Corpus of Human-Annotated Machine Translations. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-14)*, pages 2679–2682, Reykjavik, Iceland, May.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 1–46, Lisbon, Portugal, September.

Federico, M., Negri, M., Bentivogli, L., and Turchi, M. (2014). Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, pages 1643–1653, Doha, Qatar, October.

Fishel, M., Bojar, O., and Popović, M. (2012). Terra: a Collection of Translation Error-Annotated Corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-12)*, pages 7–14, Istanbul, Turkey, May.

Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Proceedings of the 8th Language Technologies Conference*, pages 89–94, Ljubljana, Slovenia, October.

Kirchhoff, K., Capurro, D., and Turner, A. (2012). Evaluating user preferences in machine translation using conjo int analysis. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT-12)*, pages 119–126, Trento, Italy, May.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit X*, pages 79–86, Phuket, Thailand, September.

Koponen, M. (2013). This translation is not too bad; an analysis of post-editor choices in a machine translation post-editing task. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, pages 1–9, Nice, France, September.

Lommel, A., Burchardt, A., Popović, M., Harris, K., Avramidis, E., and Uszkoreit, H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT-14)*, pages 165–172, Dubrovnik, Croatia.

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

Popović, M. and Arčan, M. (2015). Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT-15)*, pages 97–104, Antalya, Turkey.

Popović, M. and Burchardt, A. (2011). From Human to Automatic Error Classification for Machine Translation Output . In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT-11)*, pages 265–272, Leuven, Belgium, May.

Popović, M., Arčan, M., Avramidis, E., Burchardt, A., and Lommel, A. R. (2015). Poor man's lemmatisation for automatic error classification. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT-15)*, pages 105–112, Antalya, Turkey.

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

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-02)*, pages 901–904, Denver, CO, September.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-12)*, pages 2214–2218, Istanbul, Turkey, May.

Tyers, F. M. and Alperen, M. (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.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 697–702, Genoa, Italy, May.

Wisniewski, G., Singh, A. K., Segal, N., and Yvon, F. (2013). Design and analysis of a large corpus of post-edited translations: quality estimation, failure analysis and the variability of post-edition. In *Proceedings of the MT Summit XIV*, pages 117–124, Nice, France, September.

Zeman, D., Fishel, M., Berka, J., and Bojar, O. (2011). Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88, October.