What can we learn about the selection mechanism for post-editing?

Maja Popović, Eleftherios Avramidis, Aljoscha Burchardt,
David Vilar, Hans Uszkoreit
DFKI / Berlin, Germany
name.surname@dfki.de

Abstract

Post-editing is an increasingly common form of human-machine cooperation for translation. One possible support for the post-editing task is offering several machine outputs to a human translator from which then can choose the most suitable one. This paper investigates the selection process for such method to get a better insight into it so that it can be optimally automatised in future work. Experiments show that only about 70% of the selected sentences are the best ranked ones, and that selection mechanism is tightly related to edit distance. Furthermore, five types of performed edit operations are analysed: correcting word form, reordering, adding missing words, deleting extra words and correcting lexical choice.

1 Motivation and related work

Machine translation (MT) has improved considerably in recent years thus gaining recognition in the translation industry. However, machine translation outputs have not yet reached the same quality as human translations. Performing the post-editing has become a common practice for improving machine translation outputs. Therefore, more and more attention is paid to various aspects of post-editing, such as (Specia, 2011). Prediction of errors in rule-based system outputs has been investigated in (Valotkaite and Asadullah, 2012) in order to facilitate the post-editing process. Analysis of edit operations has been carried out in (Koponen, 2012) in order to understand discrepances between edit distance and translation quality (i.e. predicted post-editing effort).

Our work explores the selection criteria applied by professional translators when several translation outputs of each source sentence are offered for post-editing. The scenario is similar to the one in (He et al., 2010), but our approach goes beyond, since they consider only two outputs (one produced by statistical machine translation system and other by translation memory), they do not examine ranking of these outputs, they have not tested their automatic method by professional translators, and they do not analyse edit distances and the performed edit operations. Our main questions are:

- Is the translation output which is best for post-editing also the best ranked one?
- Is the edit distance of the chosen output lower than edit distances of the other outputs?
- Are there some (less) preferred edit operations?

and to the best of our knowledge they have not been investigated yet.

2 Experimental setup

The translation outputs investigated in this work are produced by German-English, German-French and German-Spanish machine translation systems in both directions. The test sets consist of three domains: news texts taken from WMT tasks (Callison-Burch et al., 2010), technical documentation extracted from the freely available OpenOffice project (Tiedemann, 2009) and client data owned by project partners. The number of

Sharon O'Brien, Michel Simard and Lucia Specia (eds.)
Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, September 2, 2013, p. 11–17.

	News	OpenOffice	Client	Total
de-en	1788	418	500	2706
de-es	514	414	548	1476
de-fr	912	412	382	1706
en-de	1744	414	0	2158
es-de	101	413	1028	1542
fr-de	1852	412	0	2264
Total	6911	2483	2458	11852

Table 1: Test sets for ranking task and selecting for post-edit task – number of source sentences per language pair and domain.

source sentences per language pair and domain can be seen in Table 4.

Four translation systems were used: a phrase-based statistical machine translation (SMT) system Moses (Koehn et al., 2007), a hierarchical SMT system Jane (Vilar et al., 2010), a commercial rule-based system Lucy (Alonso and Thurmair, 2003), and another commercial rule-based system RBMT[1].

The translation outputs generated by the described systems were then given to professional translators in order to perform ranking and post-editing using the browser-based evaluation tool Appraise (Federmann, 2010).

Ranking and post-editing tasks were defined as follows:

Ranking: for each source sentence (11852 sentences in total), rank the outputs of four different MT systems according to *how well these preserve the meaning of the source sentence*. Ties were allowed.

Select and post-edit: for each source sentence (11852 sentences in total), select the translation output *which is easiest to post-edit* and perform the editing.

Post-edit all: for each source sentence in the selected subset (4070 sentences in total), post-edit all four produced translation outputs.

For both post-editing tasks, the translators were asked to perform only the minimal post-editing necessary to achieve acceptable translation quality. Post-editing all translation outputs is a more complex and time-consuming task in comparison to post-editing only the selected outputs, therefore only a subset of source sentences was selected.

[1]The system's name is not mentioned here by request of the vendor.

rank	1	2	3	4
Overall	71.7	19.1	6.5	2.7
News	70.0	20.4	7.2	2.3
OpenOffice	62.3	24.4	8.0	5.2
Client	84.1	10.4	3.6	1.7
de-en	69.4	20.1	7.0	3.5
de-es	80.4	15.0	3.8	0.8
de-fr	68.0	21.1	8.1	2.8
en-de	66.3	22.1	8.9	2.7
es-de	77.4	15.5	3.8	3.3
fr-de	67.4	21.1	7.8	3.6

Table 2: Percentage of sentences with a given rank selected as the best for post-editing.

3 Results

3.1 Selection vs. ranking

The first question we want to answer is how the sentences chosen for post-editing were ranked in the ranking task. Table 2 shows the percentage of selected sentences for each of four ranks (1 being the best, 4 the worst). It can be seen that overall, only 70% of selected sentences were ranked as best. About 20% of selected sentences were ranked as second best, and 10% had one of the two lowest ranks. For the client data, the percentage of the first ranked selected sentences is higher (84%) as well as for the language pair German–Spanish in both translation directions, and for the technical documentation is lower (62%). The results for the rest of domains and language pairs show the same tendency as the overall results.

Table 3 shows an example of a third ranked translation selected for post-editing extracted from German-to-English client data: one word remained untranslated which degraded significantly the quality. On the other hand, the correction of this sentence is easy – it requires only one edit operation, namely replacing this (German) word with the correct (English) one. This shows that the post-editor's expectations about the amount of editing necessary, which could be approximated by the edit distance, are taken into account when it comes to select the translation to be post-edited.

source	Dazu ist ein Schraubendreher erforderlich.
Rank	Translation output
1	For this purpose a screwdriver is necessary.
2	In addition a screwdriver is necessary.
3*	This requires a Schraubendreher.
4	This would require an Schraubendreher required.
edit(3)	This requires a screwdriver.

Table 3: Example of discrepance between ranking and post-editing: the third ranked sentence is chosen for post-editing due to lower edit distance.

	edit rates (%) selected	rest	relative difference (%)
form	2.9	4.5	36.2
order	5.3	7.8	31.9
missing	3.6	6.7	**45.8**
extra	6.0	9.0	34.2
lexical	21.2	33.0	35.8
total	39.0	61.0	36.0

Table 4: Total edit distance and five distinct types of edits (%) for selected sentences and not selected sentences (first row) and their relative differences (%) (second row).

3.2 Edit distances

The previous results show that there is a difference between the selection mechanisms for ranking translation outputs based on meaning and for choosing the output most suitable for post-editing. The results also confirmed that the edit distance plays an important role for the post-editing selection, but the further question is how exactly. It would be good to know if only the total edit distance matters, or some types of edit operations are more or less preferred than the others.

In order to explore these aspects, automatic edit analysis was carried out using the Hjerson tool (Popović, 2011) using the post-edited translations as references. The following five types of edit operations were distinguished: correcting word form (morphology), correcting word order, adding missing word, deleting extra word and correcting lexical choice. The results are presented in the form of edit rates, i.e. the total number of edit operations normalised over the total number of words. The total edit distance was calculated as a sum of the five edit rates.

3.2.1 Selected vs. rest

The first step in edit distance analysis was to compare edit distances of the selected sentences with the edit distance of the remaining sentences which were not selected. The obtained edit rates together with the relative differences $(\mathrm{editRate}(rest) - \mathrm{editRate}(sel))/\mathrm{editRate}(rest)$ are presented in Table 4. The first two columns show the edit rates for selected sentences and for the rest, and the third column presents their relative differences. Overall, the relative difference between the edit distances of two sets is 36%, meaning that 36% less edit operations were performed in the selected sentences than in the rest of the sentences. The relative differences are similar for all edit operation types being between 30% and 36%, except the missing words with 45% – adding missing words does not seem to be preferred in general. Further analysis is necessary for drawing definite conclusions.

We carried out a further analysis in a somewhat different direction, namely compare the selected sentences which are not best ranked with their best ranked ”opponents”.

3.2.2 Selected vs. best ranked

Further analysis was constrained only on the best ranked sentences which were not selected for post-editing. The first step was to calculate total edit distances of these sentences and their selected counterparts, and the results are presented in Table 5. Overall, the edit distance for the selected sentences is lower than for the best ranked confirming that the edit distance is a very important factor for the selection mechanism. Separate edit distances for three distinct domains show the same tendencies. However, the results for separated translation directions showed that there are exceptions – some selected sentences have higher edit distance than their best ranked “opponents”. Two examples from the German-to-English task are shown in Table 6 and further analysis of such sentences is shown in the next section.

The first example shows a preference to two lexical corrections over one reordering. In the sec-

edit distance (%)	selected	rank 1
Overall	37.1	48.9
News	38.5	48.0
OpenOffice	33.3	51.3
Client	39.5	44.7
de-en	30.8	38.9
de-es*	**35.9**	33.9
de-fr	57.8	67.8
en-de*	**44.9**	37.6
es-de	32.8	51.7
fr-de	42.4	44.5

Table 5: Total edit distances (%) for selected sentences and best ranked not selected sentences: values are normalised over the total number of words.

source	Inzwischen sei das träge fliessende Gewässer vollkommen tot .
rank 1	Now $\mathbf{are}_{reord}$ the lazy river waters completely dead .
edit	Now the lazy river waters are completely dead .
selected (rank 3)	Meanwhile the $\mathbf{sluggish}_{lex}$ $\mathbf{river}_{lex}$ waters are completely dead .
edit	Meanwhile the sluggishly flowing waters are completely dead .
source	Probleme gibt es auch bei Kilometer 185 der Autobahn D1 in Richtung Prag .
rank 1	There are $\mathbf{problems}_{reord}$ $\mathbf{also}_{reord}$ at kilometer 185 of the motorway D1 in direction Prague .
edit	There are also problems at kilometer 185 of the motorway D1 in the direction $\mathbf{of}_{miss}$ Prague .
selected (rank 2)	There are $\mathbf{problems}_{reord}$ $\mathbf{also}_{reord}$ $\mathbf{with}_{lex}$ kilometer $\mathbf{of}_{reord}$ $\mathbf{185}_{reord}$ the motorway D1 toward Prague .
edit	There are also problems at kilometer 185 of the motorway D1 toward Prague .

Table 6: Examples of not best ranked selected sentences with larger edit distance than their best ranked counterparts.

ond example, two reorderings and one lexical corrections are performed in the selected sentence whereas in the best ranked one there are only one reordering and one omission, suggesting again that translators tend to avoid adding missing words.

3.2.3 Selected vs. best ranked with lower edit distance

Comparison between edit distances of not best ranked selected sentences and their best ranked "opponents" in the previous section generated a new question: why the translators sometimes choose sentences which are neither the best translations nor the closest translations? Further edit rate analysis was constrained only on those sentences, namely the selected sentences which are neither best ranked nor have the lowest edit distance and the five edit rates for those sentences are graphically presented in Figures 1 and 2.

Figure 1 presents the overall five edit rates together with the edit rates for each of three domains. As expected, all edit rates are higher for the selected sentences, and furthermore it can be noted that the differences are largest for reordering edit rates. Only for the technical (OpenOffice) all differences are very small.

The results for different translation directions are shown in Figure 2, and it can be seen that the differences between edit rates are rather language-dependent, although a larger reordering edit rate for selected sentences can be observed for all translation directions. On the other hand, word form (inflection) edit rate of selected sentences is significantly higher only for the English-to-German translation. A possible explanation is that the German inflections often cannot be generated properly when translating from morphologically poorer English, however correcting them does not pose a big problem for translators, especially in comparison to other edit operations. Another interesting observation is that there are neither significant nor conclusive differences between the effects of missing words – a thorough analysis of this edit operation should be carried out, however it seems that although adding missing words is generally not the preferred action for the translators, it does not influence significantly the selection of a low(er) ranked sentence.

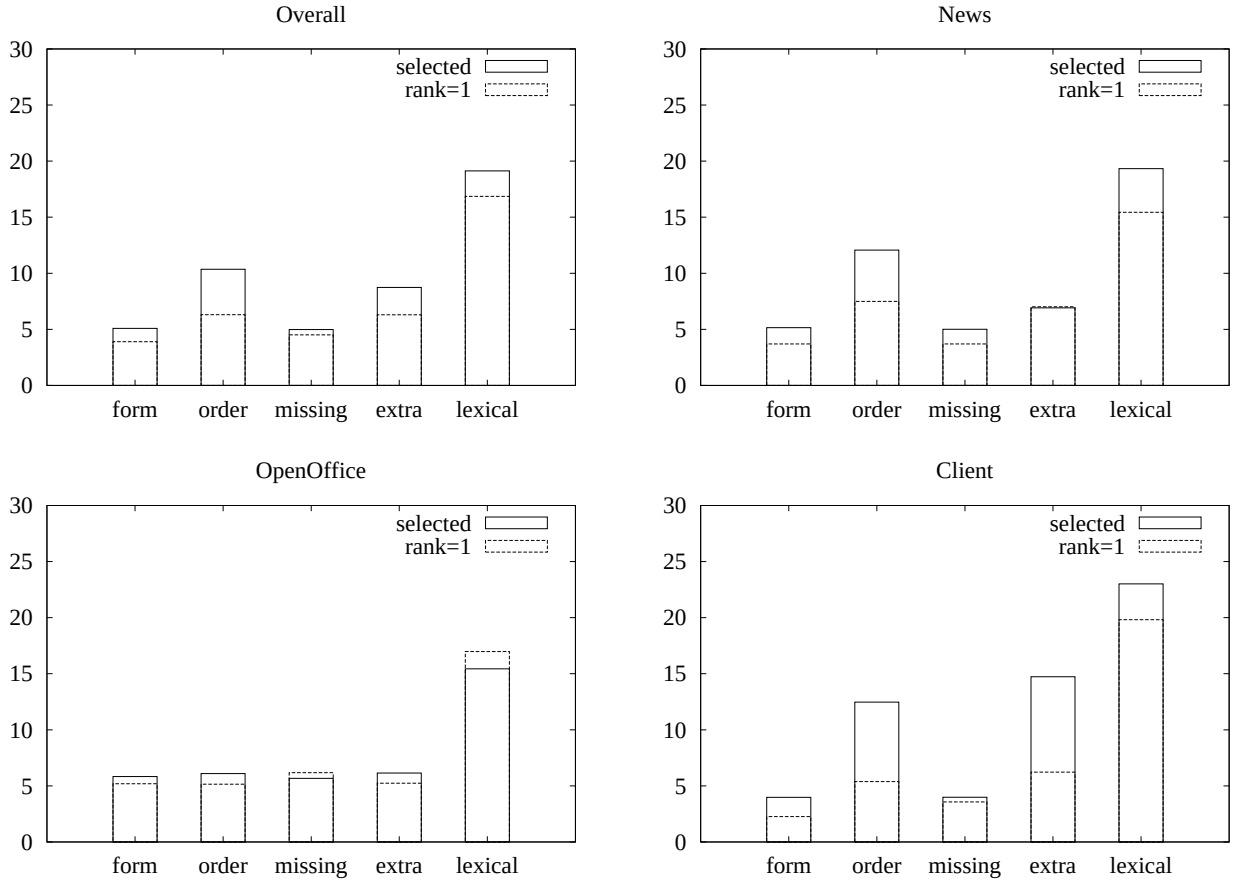

Figure 1: Five edit rates (%) of selected sentences with lower rank and larger edit distance (selected) and their best ranked counterparts (rank 1): overall and for three different domains separately.

Further analysis for different translation directions is carried out in the form of the distribution of edit operations over POS tags. For all language pairs, reordering of the noun is performed much more often in the selected sentences than in others. In addition, the amount of preposition reordering edit operations differs for all translations from German, whereas inflection and reordering of determiners are distinctive in all translations into German.

4 Summary and outlook

In this paper we investigated the post-editing selection mechanism of human translators by analysis of ranks, total edit distances and five types of edit operations. It is shown that only about 70% of the selected sentences are at the same time the best ranked ones, therefore the selection mechanisms for the best output and for the output best for post-editing differ significantly. Furthermore, it is shown that the post-editing selection mechanism may be modelled in terms of the post-editor's perception of the amount of post-editing needed, which may be measured a posteriori using the actual edit distance between the raw and the post-edited sentence. Nevertheless, a simple edit distance is not the only criterion. Further analysis has shown that some phenomena are rather language-dependent, however reordering edit operation is distinctive for all test sets. In addition, it is shown that reordering of nouns plays a significant role for all translation directions.

This work can be extended in various ways. One direction is using the obtained results for already mentioned automatisation of the selection process. Another direction is investigation of selection criteria for different translation systems, e.g. comparing statistical and rule-based systems. Furthermore, more detailed analysis including distinct types of edit operations and POS tags as well as

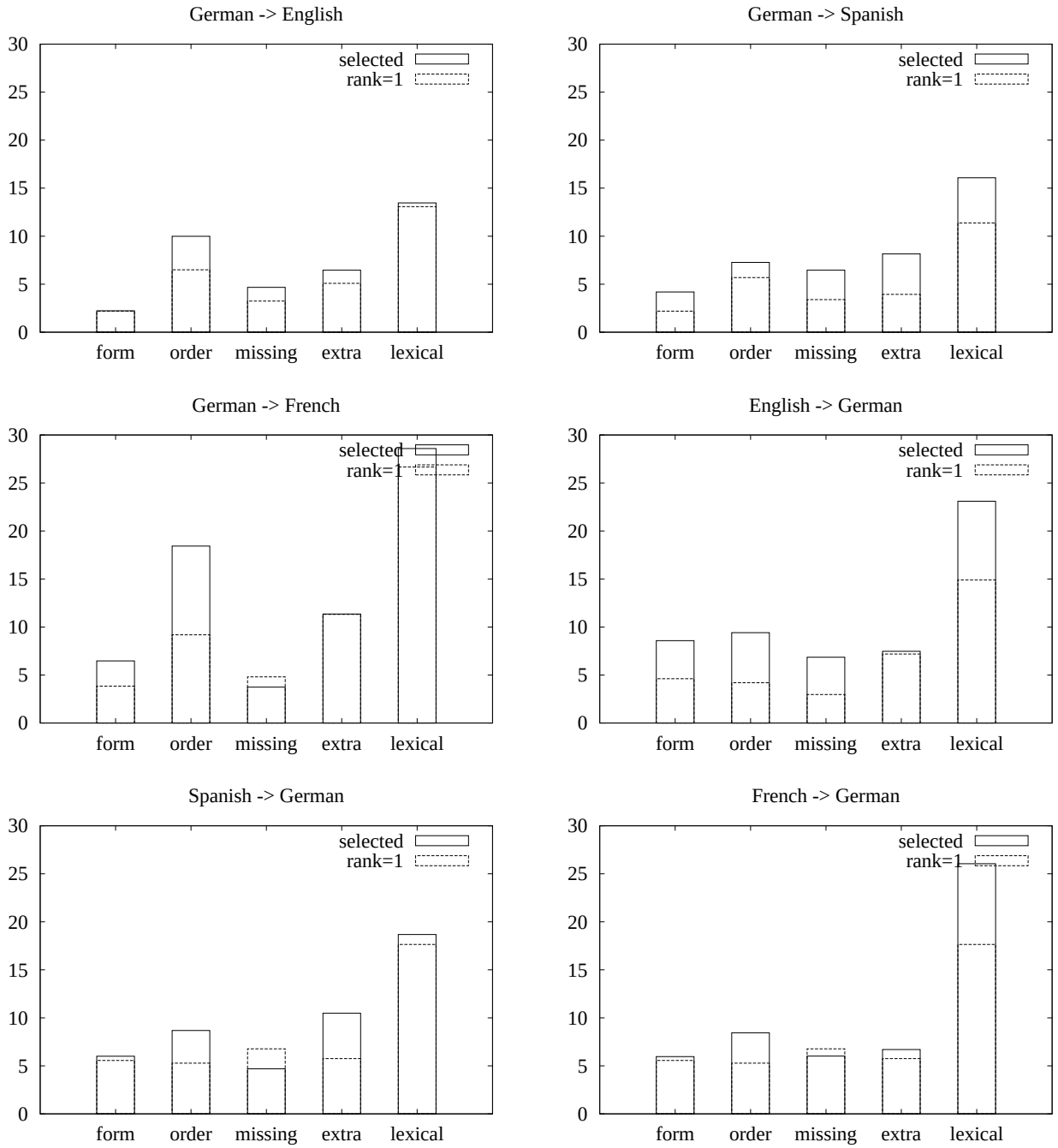

Figure 2: Five edit rates (%) of selected sentences with lower rank and larger edit distance (selected) and their best ranked counterparts (rank 1): the six translation directions are shown separately.

further investigation of missing words in various scenarios should be carried out on different language pairs and translation directions.

Acknowledgments

This work has been developed within the TARAXŰ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

References

Alonso, Juan A. and Gregor Thurmair. 2003. The comprendium translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, September.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT 2010)*, pages 17–53, Uppsala, Sweden, July.

Federmann, Christian. 2010. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May.

He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 622–630, Uppsala, Sweden, July.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Chris Zens, Richard a nd 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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190, Montral, Canada, June. Association for Computational Linguistics.

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.

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 73–80, Leuven, Belgium, May.

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

Valotkaite, Justina and Munshi Asadullah. 2012. Error Detection for Post-editing Rule-based Machine Translation. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, page 7886, San Diego, USA, October. Association for Machine Translation in the Americas (AMTA).

Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 262–270, Uppsala, Sweden, July.