Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis

Maja Popović
Lehrstuhl für Informatik 6
RWTH Aachen University
Aachen, Germany
popovic@cs.rwth-aachen.de

Hermann Ney
Lehrstuhl für Informatik 6
RWTH Aachen University
Aachen, Germany
ney@cs.rwth-aachen.de

Abstract

Evaluation and error analysis of machine translation output are important but difficult tasks. In this work, we propose a novel method for obtaining more details about actual translation errors in the generated output by introducing the decomposition of Word Error Rate (WER) and Position independent word Error Rate (PER) over different Part-of-Speech (POS) classes. Furthermore, we investigate two possible aspects of the use of these decompositions for automatic error analysis: estimation of inflectional errors and distribution of missing words over POS classes. The obtained results are shown to correspond to the results of a human error analysis. The results obtained on the European Parliament Plenary Session corpus in Spanish and English give a better overview of the nature of translation errors as well as ideas of where to put efforts for possible improvements of the translation system.

1 Introduction

Evaluation of machine translation output is a very important but difficult task. Human evaluation is expensive and time consuming. Therefore a variety of automatic evaluation measures have been studied over the last years. The most widely used are Word Error Rate (WER), Position independent word Error Rate (PER), the BLEU score (Papineni et al., 2002) and the NIST score (Doddington, 2002). These measures have shown to be valuable tools for comparing different systems as well as for evaluating improvements within one system. However, these measures do not give any details about the nature of translation errors. Therefore some more detailed analysis of the generated output is needed in order to identify the main problems and to focus the research efforts. A framework for human error analysis has been proposed in (Vilar et al., 2006), but as every human evaluation, this is also a time consuming task.

This article presents a framework for calculating the decomposition of WER and PER over different POS classes, i.e. for estimating the contribution of each POS class to the overall word error rate. Although this work focuses on POS classes, the method can be easily extended to other types of linguistic information. In addition, two methods for error analysis using the WER and PER decompositons together with base forms are proposed: estimation of inflectional errors and distribution of missing words over POS classes. The translation corpus used for our error analysis is built in the framework of the TC-STAR project (tcs, 2005) and contains the transcriptions of the European Parliament Plenary Sessions (EPPS) in Spanish and English. The translation system used is the phrase-based statistical machine translation system described in (Vilar et al., 2005; Matusov et al., 2006).

2 Related Work

Automatic evaluation measures for machine translation output are receiving more and more attention in the last years. The BLEU metric (Papineni et al., 2002) and the closely related NIST metric (Doddington, 2002) along with WER and PER

Proceedings of the Second Workshop on Statistical Machine Translation, pages 48–55,
Prague, June 2007.

have been widely used by many machine translation researchers. An extended version of BLEU which uses n-grams weighted according to their frequency estimated from a monolingual corpus is proposed in (Babych and Hartley, 2004). (Leusch et al., 2005) investigate preprocessing and normalisation methods for improving the evaluation using the standard measures WER, PER, BLEU and NIST. The same set of measures is examined in (Matusov et al., 2005) in combination with automatic sentence segmentation in order to enable evaluation of translation output without sentence boundaries (e.g. translation of speech recognition output). A new automatic metric METEOR (Banerjee and Lavie, 2005) uses stems and synonyms of the words. This measure counts the number of exact word matches between the output and the reference. In a second step, unmatched words are converted into stems or synonyms and then matched. The TER metric (Snover et al., 2006) measures the amount of editing that a human would have to perform to change the system output so that it exactly matches the reference. The CDER measure (Leusch et al., 2006) is based on edit distance, such as the well-known WER, but allows reordering of blocks. Nevertheless, none of these measures or extensions takes into account linguistic knowledge about actual translation errors, for example what is the contribution of verbs in the overall error rate, how many full forms are wrong whereas their base forms are correct, etc. A framework for human error analysis has been proposed in (Vilar et al., 2006) and a detailed analysis of the obtained results has been carried out. However, human error analysis, like any human evaluation, is a time consuming task.

Whereas the use of linguistic knowledge for improving the performance of a statistical machine translation system is investigated in many publications for various language pairs (like for example (Nießen and Ney, 2000), (Goldwater and McClosky, 2005)), its use for the analysis of translation errors is still a rather unexplored area. Some automatic methods for error analysis using base forms and POS tags are proposed in (Popović et al., 2006; Popović and Ney, 2006). These measures are based on differences between WER and PER which are calculated separately for each POS class using subsets extracted from the original texts. Standard overall WER and PER of the original texts are not at all taken into account. In this work, the standard WER and PER are decomposed and analysed.

3 Decomposition of WER and PER over POS classes

The standard procedure for evaluating machine translation output is done by comparing the hypothesis document hyp with given reference translations ref, each one consisting of K sentences (or segments). The reference document ref consists of R reference translations for each sentence. Let the length of the hypothesis sentence hyp_k be denoted as N_{hyp_k}, and the reference lengths of each sentence $N_{ref_{k,r}}$. Then, the total hypothesis length of the document is $N_{hyp} = \sum_k N_{hyp_k}$, and the total reference length is $N_{ref} = \sum_k N^*_{ref_k}$ where $N^*_{ref_k}$ is defined as the length of the reference sentence with the lowest sentence-level error rate as shown to be optimal in (Leusch et al., 2005).

3.1 Standard word error rates (overview)

The word error rate (WER) is based on the Levenshtein distance (Levenshtein, 1966) - the minimum number of substitutions, deletions and insertions that have to be performed to convert the generated text hyp into the reference text ref. A shortcoming of the WER is the fact that it does not allow reorderings of words, whereas the word order of the hypothesis can be different from word order of the reference even though it is correct translation. In order to overcome this problem, the position independent word error rate (PER) compares the words in the two sentences without taking the word order into account. The PER is always lower than or equal to the WER. On the other hand, shortcoming of the PER is the fact that the word order can be important in some cases. Therefore the best solution is to calculate both word error rates.

Calculation of WER: The WER of the hypothesis hyp with respect to the reference ref is calculated as:

$$\text{WER} = \frac{1}{N^*_{ref}} \sum_{k=1}^{K} \min_r d_L(ref_{k,r}, hyp_k)$$

where $d_L(ref_{k,r}, hyp_k)$ is the Levenshtein distance between the reference sentence $ref_{k,r}$ and the hypothesis sentence hyp_k. The calculation of WER is performed using a dynamic programming algorithm.

Calculation of PER: The PER can be calculated using the counts $n(e, hyp_k)$ and $n(e, ref_{k,r})$ of a word e in the hypothesis sentence hyp_k and the reference sentence $ref_{k,r}$ respectively:

$$\text{PER} = \frac{1}{N^*_{ref}} \sum_{k=1}^{K} \min_r d_{\text{PER}}(ref_{k,r}, hyp_k)$$

where

$$d_{\text{PER}}(ref_{k,r}, hyp_k) = \frac{1}{2}\Big(|N_{ref_{k,r}} - N_{hyp_k}| + \sum_e |n(e, ref_{k,r}) - n(e, hyp_k)| \Big)$$

3.2 WER decomposition over POS classes

The dynamic programming algorithm for WER enables a simple and straightforward identification of each erroneous word which actually contributes to WER. Let err_k denote the set of erroneous words in sentence k with respect to the best reference and p be a POS class. Then $n(p, err_k)$ is the number of errors in err_k produced by words with POS class p. It should be noted that for the substitution errors, the POS class of the involved reference word is taken into account. POS tags of the reference words are also used for the deletion errors, and for the insertion errors the POS class of the hypothesis word is taken. The WER for the word class p can be calculated as:

$$\text{WER}(p) = \frac{1}{N^*_{ref}} \sum_{k=1}^{K} n(p, err_k)$$

The sum over all classes is equal to the standard overall WER.

An example of a reference sentence and hypothesis sentence along with the corresponding POS tags is shown in Table 1. The WER errors, i.e. actual words participating in WER together with their POS classes can be seen in Table 2. The reference words involved in WER are denoted as reference errors, and hypothesis errors refer to the hypothesis words participating in WER.

Standard WER of the whole sentence is equal to $4/12 = 33.3\%$. The contribution of nouns is $\text{WER(N)} = 1/12 = 8.3\%$, of verbs $\text{WER(V)} = 2/12 = 16.7\%$ and of adverbs $\text{WER(ADV)} = 1/12 = 8.3\%$

reference: Mister#N Commissioner#N ,#PUN twenty-four#NUM hours#N sometimes#ADV can#V be#V too#ADV much#PRON time#N .#PUN
hypothesis: Mrs#N Commissioner#N ,#PUN twenty-four#NUM hours#N is#V sometimes#ADV too#ADV much#PRON time#N .#PUN

Table 1: Example for illustration of actual errors: a POS tagged reference sentence and a corresponding hypothesis sentence

reference errors	hypothesis errors	error type
Mister#N	Mrs#N	substitution
sometimes#ADV	is#V	substitution
can#V		deletion
be#V	sometimes#ADV	substitution

Table 2: WER errors: actual words which are participating in the word error rate and their corresponding POS classes

3.3 PER decomposition over POS classes

In contrast to WER, standard efficient algorithms for the calculation of PER do not give precise information about contributing words. However, it is possible to identify all words in the hypothesis which do not have a counterpart in the reference, and vice versa. These words will be referred to as PER errors.

reference errors	hypothesis errors
Mister#N	Mrs#N
be#V	is#V
can#V	

Table 3: PER errors: actual words which are participating in the position independent word error rate and their corresponding POS classes

An illustration of PER errors is given in Table 3.

The number of errors contributing to the standard PER according to the algorithm described in 3.1 is 3 - there are two substitutions and one deletion. The problem with standard PER is that it is not possible to detect which words are the deletion errors, which are the insertion errors, and which words are the substitution errors. Therefore we introduce an alternative PER based measure which corresponds to the F-measure. Let $herr_k$ refer to the set of words in the hypothesis sentence k which do not appear in the reference sentence k (referred to as hypothesis errors). Analogously, let $rerr_k$ denote the set of words in the reference sentence k which do not appear in the hypothesis sentence k (referred to as reference errors). Then the following measures can be calculated:

- reference PER (RPER) (similar to recall):

$$\mathrm{RPER}(p) = \frac{1}{N^*_{ref}} \sum_{k=1}^{K} n(p, rerr_k)$$

- hypothesis PER (HPER) (similar to precision):

$$\mathrm{HPER}(p) = \frac{1}{N_{hyp}} \sum_{k=1}^{K} n(p, herr_k)$$

- F-based PER (FPER):

$$\mathrm{FPER}(p) = \frac{1}{N^*_{ref} + N_{hyp}} \cdot \\ \cdot \sum_{k=1}^{K} (n(p, rerr_k) + n(p, herr_k))$$

Since we are basically interested in all words without a counterpart, both in the reference and in the hypothesis, this work will be focused on FPER. The sum of FPER over all POS classes is equal to the overall FPER, and the latter is always less or equal to the standard PER.

For the example sentence presented in Table 1, the number of hypothesis errors $n(e, herr_k)$ is 2 and the number of reference errors $n(e, rerr_k)$ is 3 where e denotes the word. The number of errors contributing to the standard PER is 3, since $|N_{ref} - N_{hyp}| = 1$ and $\sum_e |n(e, ref_k) - n(e, hyp_k)| = 5$. The standard PER is normalised over the reference length $N_{ref} = 12$ thus being equal to 25%. The FPER is the sum of hypothesis and reference errors divided by the sum of hypothesis and reference length: $\mathrm{FPER} = (2+3)/(11+12) = 5/23 = 21.7\%$. The contribution of nouns is $\mathrm{FPER(N)} = 2/23 = 8.7\%$ and the contribution of verbs is $\mathrm{FPER(V)} = 3/23 = 13\%$.

4 Applications for error analysis

The decomposed error rates described in Section 3.2 and Section 3.3 contain more details than the standard error rates. However, for more precise information about certain phenomena some kind of further analysis is required. In this work, we investigate two possible aspects for error analysis:

- estimation of inflectional errors by the use of FPER errors and base forms
- extracting the distribution of missing words over POS classes using WER errors, FPER errors and base forms.

4.1 Inflectional errors

Inflectional errors can be estimated using FPER errors and base forms. From each reference-hypothesis sentence pair, only erroneous words which have the common base forms are taken into account. The inflectional error rate of each POS class is then calculated in the same way as FPER. For example, from the PER errors presented in Table 3, the words “is” and “be” are candidates for an inflectional error because they are sharing the same base form “be”. Inflectional error rate in this example is present only for the verbs, and is calculated in the same way as FPER, i.e. $\mathrm{iFPER(V)} = 2/23 = 8.7\%$.

4.2 Missing words

Distribution of missing words over POS classes can be extracted from the WER and FPER errors in the following way: the words considered as missing are those which occur as deletions in WER errors and at the same time occur only as reference PER errors without sharing the base form with any hypothesis error. The use of both WER and PER errors is much more reliable than using only the WER deletion erros because not all deletion errors are produced by missing words: a number of WER deletions appears due to reordering errors. The information about the base form is used in order to eliminate inflectional errors. The number of missing words is extracted for each word class and then normalised over the sum of all classes. For the example sentence pair presented in Table 1, from the WER errors in Table 2 and the PER errors in Table 3 the word "can" will be identified as missing.

5 Experimental settings

5.1 Translation System

The machine translation system used in this work is based on the statistical aproach. It is built as a log-linear combination of seven different statistical models: phrase based models in both directions, IBM1 models at the phrase level in both directions, as well as target language model, phrase penalty and length penalty are used. A detailed description of the system can be found in (Vilar et al., 2005; Matusov et al., 2006).

5.2 Task and corpus

The corpus analysed in this work is built in the framework of the TC-STAR project. The training corpus contains more than one million sentences and about 35 million running words of the European Parliament Plenary Sessions (EPPS) in Spanish and English. The test corpus contains about 1 000 sentences and 28 000 running words. The OOV rates are low, about 0.5% of the running words for Spanish and 0.2% for English. The corpus statistics can be seen in Table 4. More details about the EPPS data can be found in (Vilar et al., 2005).

TRAIN	Spanish	English
Sentences	1 167 627	
Running words	35 320 646	33 945 468
Vocabulary	159 080	110 636
TEST		
Sentences	894	1 117
Running words	28 591	28 492
OOVs	0.52%	0.25%

Table 4: Statistics of the training and test corpora of the TC-STAR EPPS Spanish-English task. Test corpus is provided with two references.

6 Error analysis

The translation is performed in both directions (Spanish to English and English to Spanish) and the error analysis is done on both the English and the Spanish output. Morpho-syntactic annotation of the English references and hypotheses is performed using the constraint grammar parser ENGCG (Voutilainen, 1995), and the Spanish texts are annotated using the FreeLing analyser (Carreras et al., 2004). In this way, all references and hypotheses are provided with POS tags and base forms. The decomposition of WER and FPER is done over the ten main POS classes: nouns (N), verbs (V), adjectives (A), adverbs (ADV), pronouns (PRON), determiners (DET), prepositions (PREP), conjunctions (CON), numerals (NUM) and punctuation marks (PUN). Inflectional error rates are also estimated for each POS class using FPER counts and base forms. Additionally, details about the verb tense and person inflections for both languages as well as about the adjective gender and person inflections for the Spanish output are extracted. Apart from that, the distribution of missing words over the ten POS classes is estimated using the WER and FPER errors.

6.1 WER and PER (FPER) decompositions

Figure 1 presents the decompositions of WER and FPER over the ten basic POS classes for both languages. The largest part of both word error rates comes from the two most important word classes, namely nouns and verbs, and that the least critical classes are punctuations, conjunctions and numbers.

Adjectives, determiners and prepositions are significantly worse in the Spanish output. This is partly due to the richer morphology of the Spanish language. Furthermore, the histograms indicate that the number of erroneus nouns and pronouns is higher in the English output. As for verbs, WER is higher for English and FPER for Spanish. This indicates that there are more problems with word order in the English output, and more problems with the correct verb or verb form in the Spanish output.

In addition, the decomposed error rates give an idea of where to put efforts for possible improvements of the system. For example, working on improvements of verb translations could reduce up to about 10% WER and 7% FPER, working on nouns

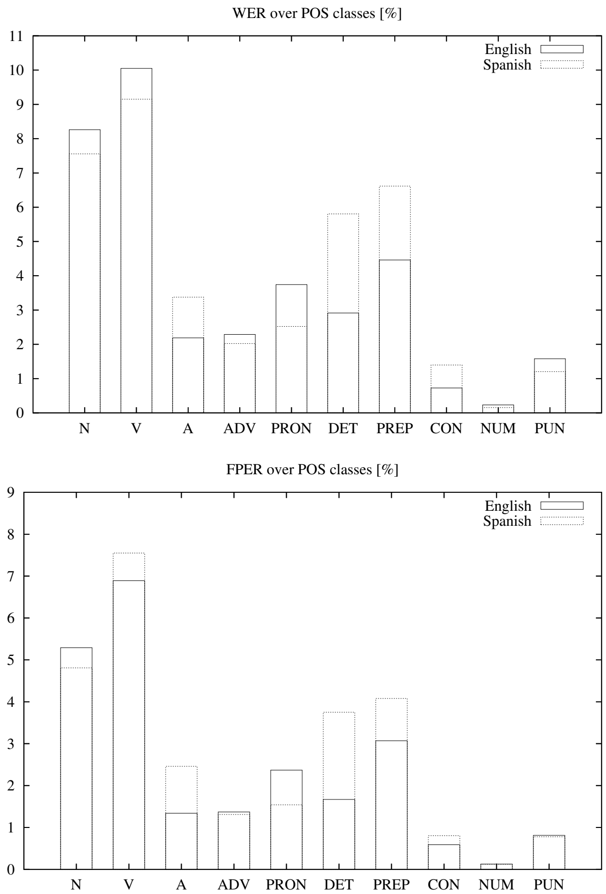

Figure 1: Decomposition of WER and FPER [%] over the ten basic POS classes for English and Spanish output

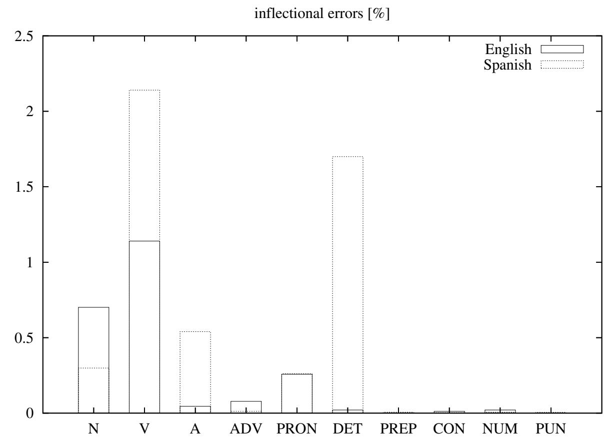

Figure 2: Inflectional error rates [%] for English and Spanish output

up to 8% WER and 5% FPER, whereas there is no reason to put too much efforts on e.g. adverbs since this could lead only to about 2% of WER and FPER reduction. [1]

6.2 Inflectional errors

Inflectional error rates for the ten POS classes are presented in Figure 2. For the English language, these errors are significant only for two POS classes: nouns and verbs. The verbs are the most problematic category in both languages, for Spanish having almost two times higher error rate than for English. This is due to the very rich morphology of Spanish verbs - one base form might have up to about fourty different inflections.

[1]Reduction of FPER leads to a similar reduction of PER.

Nouns have a higher error rate for English than for Spanish. The reason for this difference is not clear, since the noun morphology of neither of the languages is particularly rich - there is only distinction between singular and plural. One possible explanation might be the numerous occurences of different variants of the same word, like for example "Mr" and "Mister".

In the Spanish output, two additional POS classes are showing significant error rate: determiners and adjectives. This is due to the gender and number inflections of those classes which do not exist in the English language - for each determiner or adjective, there are four variants in Spanish and only one in English. Working on inflections of Spanish verbs might reduce approximately 2% of FPER, on English verbs about 1%. Improvements of Spanish determiners could lead up to about 2% of improvements.

6.2.1 Comparison with human error analysis

The results obtained for inflectional errors are comparable with the results of a human error analysis carried out in (Vilar et al., 2006). Although it is difficult to compare all the numbers directly, the overall tendencies are the same: the largest number of translation errors are caused by Spanish verbs, and much less but still a large number of errors by English verbs. A much smaller but still significant number of errors is due to Spanish adjectives, and only a few errors of English adjectives are present.

Human analysis was done also for the tense and person of verbs, as well as for the number and gender of adjectives. We use more detailed POS tags in order to extract this additional information and calculate inflectional error rates for such tags. It should be noted that in contrast to all previous error rates, these error rates are not disjunct but overlapping: many words are contributing to both.

The results are shown in Figure 3, and the tendencies are again the same as those reported in (Vilar et al., 2006). As for verbs, tense errors are much more frequent than person errors for both languages. Adjective inflections cause certain amount of errors only in the Spanish output. Contributions of gender and of number are aproximately equal.

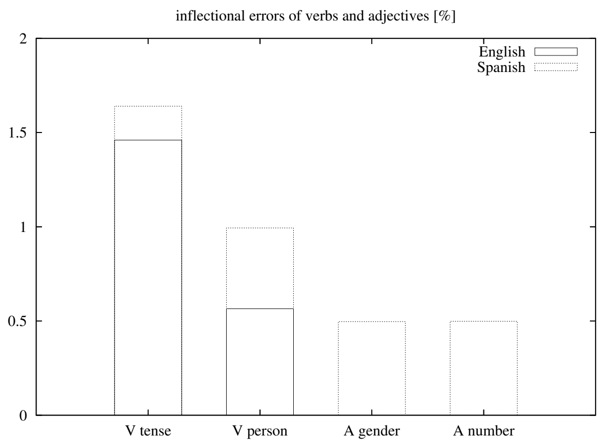

Figure 3: More details about inflections: verb tense and person error rates and adjective gender and number error rates [%]

6.3 Missing words

Figure 4 presents the distribution of missing words over POS classes. This distribution has a same behaviour as the one obtained by human error analysis. Most missing words for both languages are verbs. For English, the percentage of missing verbs is significantly higher than for Spanish. The same thing happens for pronouns. The probable reason for this is the nature of Spanish verbs. Since person and tense are contained in the suffix, Spanish pronouns are often omitted, and auxiliary verbs do not exist for all tenses. This could be problematic for a translation system, because it processes only one Spanish word which actually contains two (or more) English words.

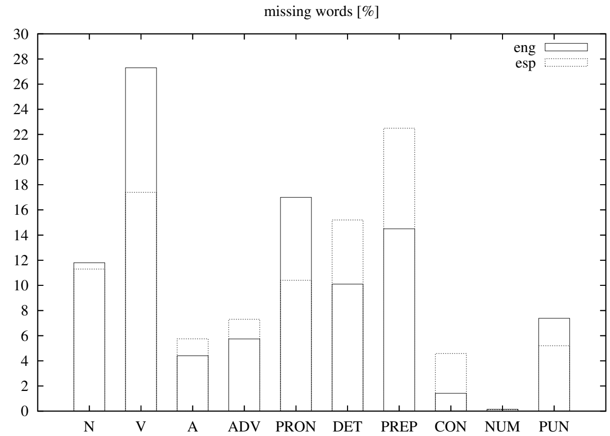

Figure 4: Distribution of missing words over POS classes [%] for English and Spanish output

Prepositions are more often missing in Spanish than in English, as well as determiners. A probable reason is the disproportion of the number of occurrences for those classes between two languages.

7 Conclusions

This work presents a framework for extraction of linguistic details from standard word error rates WER and PER and their use for an automatic error analysis. We presented a method for the decomposition of standard word error rates WER and PER over ten basic POS classes. We also carried out a detailed analysis of inflectional errors which has shown that the results obtained by our method correspond to those obtained by a human error analysis. In addition, we proposed a method for analysing missing word errors.

We plan to extend the proposed methods in order to carry out a more detailed error analysis, for example examining different types of verb inflections. We also plan to examine other types of translation errors like for example errors caused by word order.

Acknowledgements

This work was partly funded by the European Union under the integrated project TC-STAR– Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

References

Bogdan Babych and Anthony Hartley. 2004. Extending BLEU MT Evaluation Method with Frequency Weighting. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, July.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, June.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC)*, pages 239–242, Lisbon, Portugal, May.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*, pages 128–132, San Diego.

Sharon Goldwater and David McClosky. 2005. Improving stastistical machine translation through morphological analysis. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Vancouver, Canada, October.

Gregor Leusch, Nicola Ueffing, David Vilar, and Hermann Ney. 2005. Preprocessing and Normalization for Automatic Evaluation of Machine Translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 17–24, Ann Arbor, MI, June. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *EACL06*, pages 241–248, Trento, Italy, April.

Vladimir Iosifovich Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, February.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 148–154, Pittsburgh, PA, October.

Evgeny Matusov, Richard Zens, David Vilar, Arne Mauser, Maja Popović, and Hermann Ney. 2006. The RWTH Machine Translation System. In *TC-Star Workshop on Speech-to-Speech Translation*, pages 31–36, Barcelona, Spain, June.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1081–1085, Saarbrücken, Germany, July.

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

Maja Popović and Hermann Ney. 2006. Error Analysis of Verb Inflections in Spanish Translation Output. In *TC-Star Workshop on Speech-to-Speech Translation*, pages 99–103, Barcelona, Spain, June.

Maja Popović, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico, and Rafael Banchs. 2006. Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. In *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 1–6, New York, NY, June.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proc. of the 7th Conf. of the Association for Machine Translation in the Americas (AMTA 06)*, pages 223–231, Boston, MA.

2005. TC-STAR - technology and corpora for speech to speech translation. Integrated project TCSTAR (IST-2002-FP6-506738) funded by the European Commission. http://www.tc-star.org/.

David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. 2005. Statistical Machine Translation of European Parliamentary Speeches. In *Proc. MT Summit X*, pages 259–266, Phuket, Thailand, September.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, pages 697–702, Genoa, Italy, May.

Atro Voutilainen. 1995. ENGCG - Constraint Grammar Parser of English. http://www2.lingsoft.fi/doc/engcg/intro/.