Readability Assessment of Translated Texts

Alina Maria Ciobanu[1,2], Liviu P. Dinu[1,2], Flaviu Ioan Pepelea[3]
[1]Faculty of Mathematics and Computer Science, University of Bucharest
[2]Center for Computational Linguistics, University of Bucharest
[3]Twitter
alina.ciobanu@my.fmi.unibuc.ro,
liviu.p.dinu@gmail.com, flaviupepelea@gmail.com

Abstract

In this paper we investigate how readability varies between texts originally written in English and texts translated into English. For quantification, we analyze several factors that are relevant in assessing readability – shallow, lexical and morpho-syntactic features – and we employ the widely used Flesch-Kincaid formula to measure the variation of the readability level between original English texts and texts translated into English. Finally, we analyze whether the readability features have enough discriminative power to distinguish between originals and translations.

1 Introduction and Related Work

The products of translation generally differ from original, non-translated texts. According to Koppel and Ordan (2011), two main aspects that lead to differences between the two categories have been identified: 1) effects of the translation process that are independent of the source language; 2) effects of the source language on the translation product, also known as source language interference. According to Sun (2012), the reception of a translated text is related to cross-cultural readability. Translators need to understand the particularities of both the source and the target language in order to transfer the meaning of the text from one language to another. While rendering the source language text into the target language, it is also important to maintain the style of the document. Various genres of text might be translated for different purposes, which influence the choice of the translation strategy. For example, for political speeches the purpose is to report exactly what is communicated in a given text (Trosborg, 1997). In this paper we investigate how readability features differ between original and translated texts.

Systems for automatic readability assessment have received an increasing attention during the last decade. While research focused initially on English, further studies have shown a growing interest in other languages, such as Spanish (Huerta, 1959), French (Kandel and Moles, 1958) or Italian (Franchina and Vacca, 1986; François and Miltsakaki, 2012). Readability assessment systems have a wide variety of applications. We mention here only a few: 1) they provide assistance in selecting reading material with an appropriate level of complexity from a large collection of documents, for second language learners and people with disabilities or low literacy skills (Collins-Thompson, 2011); 2) they help adapting the technical documents to various levels of medical expertise, within the medical domain (Elhadad and Sutaria, 2007); 3) they assist the processes of machine translation, text simplification, or speech recognition and evaluate their effectiveness, in the research area of NLP (Aluisio et al., 2010; Stymne et al., 2013).

Most of the traditional readability approaches investigate shallow text properties to determine the complexity of a text, based on assumptions which correlate surface features with the linguistic factors which influence readability. For example, the average number of characters or syllables per word, the average number of words per sentence and the percentage of words not occurring among the most frequent *n* words in a language are correlated with the lexical, syntactic and, respectively, the semantic complexity of the text. The Flesch-Kincaid measure (Kincaid et al., 1975) employs the average number of syllables per word and the average number of words per sentence to assess readability, while the Automated Readability Index (Smith and Senter, 1967) and the Coleman-Liau metric (Coleman and Liau, 1975) measure word length based on character count rather than syllable count; they are func-

Proceedings of Recent Advances in Natural Language Processing, pages 97–103,
Hissar, Bulgaria, Sep 7–9 2015.

tions of both the average number of characters per word and the average number of words per sentence. Gunning Fog (Gunning, 1952) and SMOG (McLaughlin, 1969) account also for the percentage of polysyllabic words and the Dale-Chall formula (Dale and Chall, 1995) relies on lists of most frequent words to assess readability.

2 Our Approach

The problem that we investigate in this paper is how the readability level varies across original and translated texts (from various source languages). We identify utterances from *Europarl* in a wide variety of languages, we identify their translations into English, and on these English translations we conduct a quantitative analysis of the readability features. As most research on readability focused on English so far, there are several formulas, features and tools available for quantifying the differences in the level of readability.

In this paper we complement our previous analysis (Ciobanu and Dinu, 2014) on the readability features for the original texts and their translations. Here we focus on the target language, analyzing whether different source languages lead to differences in the readability level for the translated texts.

2.1 Data

We run our experiments on *Europarl* (Koehn, 2005), a multilingual parallel corpus extracted from the proceedings of the European Parliament. Its main intended use is as aid for statistical machine translation research (Tiedemann, 2012). The corpus is tokenized and aligned in 21 languages. In Table 1 we report statistics extracted from our dataset. Given the fact that the Flesch-Kincaid formula is based on the average number of words per sentence and on the average number of syllables per word, the differences between the languages (in terms of the number of speakers and sentences) do not affect the results.

According to van Halteren (2008), translations in the European Parliament are generally made by native speakers of the target language. Translation is an inherent part of the political activity (Schäffner and Bassnett, 2010) and has a high influence on the way the political speeches are perceived. The question posed by Schäffner and Bassnett (2010) *"What exactly happens in the complex processes of recontextualisation across linguistic, cultural and ideological boundaries?"* summarizes the complexity of the process of translating political documents. Political texts might contain complex technical terms and elaborated sentences. Therefore, the results of our experiments are probably domain-specific and cannot be generalized to other types of text. Although parliamentary documents probably have a low readability level, our investigation is not negatively influenced by the choice of corpus because we are consistent across all experiments in terms of text gender and we report results obtained solely by comparison between source and target languages.

Lang.	# speakers	# sentences
EN	62	1,262
SV	292	80,171
NL	226	156,836
DA	151	37,045
FI	99	36,768
DE	539	300,672
ET	22	4,284
MT	15	2,790
PL	175	62,479
FR	691	264,460
LV	30	4,652
SL	41	8,576
HU	89	23,129
CS	67	20,637
BG	33	5,432
SK	35	13,873
LT	48	14,834
ES	378	116,834
RO	75	24,586
IT	389	109,297
PT	166	98,653

Table 1: Number of speakers and sentences for each language in our *Europarl* subset.

2.2 Pre-processing

To obtain the dataset for our experiments, we follow the pre-processing steps described by Ciobanu and Dinu (2014). We extract segments of text written in English, we identify their source languages, and we group them based on the language of the speaker. We compute the Flesch-Kicaid formula for each collection of segments of text T_i having the source language L_i and the target language English. The files contain annotations for marking the document (<*chapter*>), the speaker (<*speaker*>) and the paragraph (<*p*>). Some documents have the attribute *language* for the *speaker* tag, which indicates the language used by the original speaker. Another way of annotating the original language is by having the language abbreviation written between parentheses at the beginning of each segment of text. However, there are segments where the language is not marked in either of the two ways. We account only for sentences for which the original language could be determined.

We handle inconsistent encodings and values generated by the automatic extraction of the information from the website of the European Parliament, such as the occurrence of more than one speaker names in the <*speaker*> tag, separated either by a comma or by the *and* conjunction, or the occurrence of a speaker's affiliation in the <*speaker*> tag, e.g., *Ana Maria Gomes (PSE)*. We discard the transcribers' descriptions of the parliamentary sessions (such as *"Applause"* or *"The President interrupted the speaker"*).

3 Experiments

In this section we describe our experiments on the variability of the readability feature values for original English texts and texts translated into English from various source languages.

3.1 Flesch-Kincaid

We employ the Flesch-Kincaid measure (Kincaid et al., 1975), which assesses readability based on the average number of syllables per word and the average number of words per sentence. The Flesch-Kincaid formula is one of the most widely used readability metrics developed for English. It assesses the level of readability accounting for the number of syllables per word (as an approximation of the difficulty of a word) and for the number of words per sentence (as estimation of the syntactic difficulty of a text). The metric is computed as follows:

$$0.39\frac{\#words}{\#sentences} + 11.8\frac{\#syllables}{\#words} - 15.59.$$

The Flesch-Kincaid formula produces values which correspond with U.S. grade levels. We apply this measure on English texts, either originally written in English or translated from other languages. To determine the number of syllable for English words, we employ CMU Pronouncing Dictionary[1], a machine-readable dictionary that contains over 125,000 words and provides information regarding their syllabication.

In order to investigate and compare the readability level for original English texts and texts translated from other languages, we complete the following experiments. In a first phase, we compute the Flesh-Kincaid metric for each language, for all the concatenated files in our Europarl subcorpus.

3.1.1 Outliers Removal

The readability of a text depends, among other things, on its author. We investigate whether the readability level characterizes certain speakers, if it varies across different utterances of the same speaker and if the readability level for a language is influenced by speakers having odd readability levels associated. For this purpose, we designed three experiments based on the same idea – identification of outliers in our dataset. Further, in order to eliminate a confounding factor, namely the individuality of the speakers, to focus on the source language of the text, we perform three stages of pruning for our dataset.

- **S1:** For each language, we account for the overall readability score computed for all documents of each speaker; based on these computed values, we determine outliers and remove them from the dataset; then, we rerun the experiments based on Flesch-Kincaid measure for the remaining speakers. In order to achieve this, we divide the dataset based on the source language of the segments of text and for each language we divide the segments of text based on the speaker. We compute the overall readability score for the utterances of each speaker and, after dividing the segments of text from the dataset based on the speakers, we compute the standard quartiles *Q1*, *Q2* and *Q3* with regard to the overall level of readability for each speaker. We use the interquartile range $IQR = Q3 - Q1$ to find outliers in data. For our experiments, we consider outliers the observations that fall below $Q1 - 1.5(IQR)$ (lower fence - LF) or above

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

$Q3 + 1.5(IQR)$ (upper fence - UF) (Sheskin, 2003). We compute the Flesch-Kincaid formula again accounting only for the speakers having the individual level of readability in $[LF, UF]$ range.

- **S2:** We repeat the previous experiment introducing a further level of granularity: we investigate outliers for each speaker by computing the Flesch-Kincaid metric individually for each document belonging to a speaker. We discard documents whose levels of readability are outliers and we compute the Flesch-Kincaid formula again accounting only for the documents having the individual level of readability in the $[LF, UF]$ range.
- **S3:** In the last experiment we consider, for each language, the readability scores of each document belonging to each speaker. We apply the same strategy as before: we detect outliers among documents and remove them from the dataset. Then we compute the Flesch-Kincaid measure again, for the concatenation of all the remaining documents after outliers removal, for each language.

3.1.2 Results

In Table 2, column 2, we report the Flesch-Kincaid values for all 21 languages. One can notice that the lowest Flesh-Kincaid value belongs to the collection of texts having English as the source language, followed by texts having Germanic source languages, texts having Slavic source languages and, finally, texts translated from Romance languages. Finno-Ugric languages represent the only family that doesn't form a cluster with regard to the Flesch-Kincaid metric value. Among the Romance languages, French is the only one that sets apart from the group, being closer to the Germanic cluster. For the outliers removal experiment we report the results in Table 2, columns 3-5. The results are very similar to those of the initial experiment, suggesting that although there are outliers in the data (in Figure 1 we represent the boxplot for the Flesch-Kincaid values for each speaker's utterances), their presence does not impact significantly the overall readability values.

	Flesch-Kincaid			
Lang.	**before removing outliers**	**after pruning**		
		S1	**S2**	**S3**
EN	11.45	11.50	11.47	11.51
SV	11.50	11.49	11.45	11.44
NL	11.56	11.55	11.51	11.50
DA	11.95	11.94	11.90	11.89
FI	11.99	12.01	11.95	11.94
DE	12.45	12.44	12.38	12.37
ET	12.71	12.71	12.66	12.62
MT	12.79	12.79	12.73	12.74
PL	12.81	12.81	12.75	12.73
FR	13.29	13.30	13.25	13.24
LV	13.34	13.34	13.25	13.26
SL	13.35	13.31	13.34	13.32
HU	13.46	13.41	13.42	13.41
CS	13.75	13.76	13.70	13.66
BG	13.90	13.73	13.80	13.84
SK	13.91	13.91	13.86	13.84
LT	14.69	14.72	14.60	14.59
ES	14.72	14.70	14.61	14.59
RO	15.01	15.00	14.91	14.88
IT	15.54	15.54	15.46	15.46
PT	15.60	15.60	15.47	15.44

Table 2: Flesch-Kincaid values for our *Europarl* subset before (column 2) and after (columns 3-5) removing outliers.

3.2 Classification

In this section we investigate the readability of translation as a classification problem. Taking as input original English sentences and sentences translated from other languages, our goal is to see whether the readability features have enough discriminative power to distinguish original from translated text. Thus, we train a logistic regression classifier[2] for a binary decision problem: original versus translation. We extract randomly from our dataset 1,000 English original sentences and 1,000 sentences translated into English[3]. We split this dataset into train and test subsets with a 3:1 ratio. We choose the optimal value for the logistic regression regularization parameter performing 3-fold cross-validation on the training set (we search over $\{10^{-3}, ..., 10^{3}\}$). Finally, we evaluate the model on the test set.

[2] We use the *scikit-learn* library (Pedregosa et al., 2011).

[3] We work with only 1,000 sentences in order to have a stratified dataset, since for English the number of sentences we identified is 1,262. The subset of translated sentences is also stratified: 50 from each of the 20 languages that we investigate, besides English.

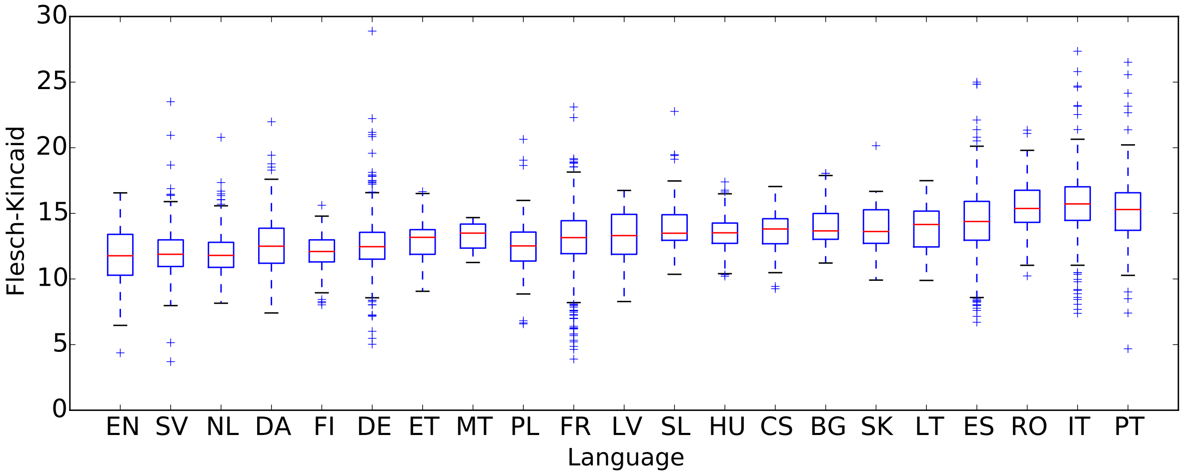

Figure 1: Boxplot for the Flesch-Kincaid values for each speaker's utterances, grouped by the language of the speaker.

3.2.1 Features

We use several shallow, lexical and morpho-syntactic features that were traditionally used for assessing readability and have proven high discriminative power within readability metrics:

- **Shallow Features**
  - **Average number of words per sentence.** The average sentence length is one of the most widely used metrics for determining readability level and was employed in numerous readability formulas, proving to be most meaningful in combined evidence with average word frequency. Feng et al. (2010) find the average sentence length to have higher predictive power than the other lexical and syllable-based features they used.
  - **Average number of characters (or syllables) per word.** It is generally considered that frequently occurring words are usually short, so the average number of characters per word was broadly used for measuring readability in a robust manner. Many readability formulas measure word length in syllables rather than letters.
- **Lexical Features**
  - **Type/Token Ratio.** The proportion between the number of lexical types and the number of tokens indicates the range of use of vocabulary. The higher the value of this feature, the higher the variability of the vocabulary used in the text.
- **Morpho-Syntactic Features**
  - **Relative frequency of POS unigrams.** The ratio for 5 POS (verbs, nouns, pronouns, adjectives and adverbs), computed individually on a per-token basis[4].
  - **Lexical density.** The proportion of content words (verbs, nouns, adjectives and adverbs), computed on a per-token basis. Grammatical features were shown to be useful in readability prediction (Heilman et al., 2007).

3.2.2 Results

The optimal value for the logistic regression regularization parameter is found to be 1. We obtain 0.59 F-score on the test set, on average, in deciding whether a sentence was translated into English or is an original English sentence. In Table 3 we report the precision, recall and F-score for the prediction task. We also report 95% confidence intervals (CI) measured on 1,000 iterations of bootstrap resampling with replacement (Koehn, 2004). The most informative features are morphological features, more specifically the POS ratios, as shown in Table 4. These results are significantly lower than state-of-the-art performance

[4] For tokenization, lemmatization and part of speech tagging we use the Stanford CoreNLP Natural Language Processing Toolkit (Manning et al., 2014).

Class	Precision	Recall	F-score
Original EN	0.60 [0.55, 0.65]	0.56 [0.51, 0.61]	0.58 [0.54, 0.62]
Translated	0.58 [0.53, 0.63]	0.62 [0.56, 0.67]	0.60 [0.56, 0.64]

Table 3: Classification results and 95% bootstrapped confidence intervals for a 2-class prediction problem — original vs. translated text — using readability features.

in translation identification, suggesting that readability features do not have enough discriminative power for the prediction task[5]. Adding n-grams of tokens and POS tags as features improves the performance of the model, leading to 0.75 average F-score ([0.71, 0.78] 95% CI) in discriminating between English sentences and translations.

4 Conclusions

In this paper we investigate the impact of translation on readability as a two-fold problem. Firstly, we investigate how the Flesch-Kincaid values vary for original English texts and for translations form different languages into English. We notice that the values form clusters for the investigated language families. Secondly, we use a set of shallow, lexical and morpho-syntactic readability features to investigate whether readability features have enough discriminative power to distinguish original English texts from translations. We obtain 0.59 F-score, on average, using only readability features, and an improvement to 0.75 when we add n-grams of tokens and POS tags as features. Our results show that, although the readability level of translated texts is similar for texts having the source language in the same language families, readability features do not have enough discriminative power to obtain high performance on distinguishing original texts from translations. However, using only readability features the prediction F-score is significantly better than chance ($p < 0.05$).

In our future work, we intend to enrich the variety of the texts, beginning with an analysis of translations of literary works. As far as resources are available, we plan to investigate other readability metrics as well. We believe our method can provide useful information regarding the difficulty of translation from one language into another in terms of readability.

Feature	Coefficient
Verb ratio	–1.59
Adverb ratio	1.49
Adjective ratio	1.35
Pronoun ratio	–1.21
Noun ratio	–0.88
Lexical density	–0.83
Type/token ratio	0.49
Average number of syllables	0.49
Average number of characters	0.04
Average number of words	–0.01

Table 4: Logistic regression coefficients for readability features (the higher the absolute value of the coefficient, the more informative the feature).

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. The contribution of the authors to this paper is equal. Liviu P. Dinu was supported by UEFISCDI, PNII-ID-PCE-2011-3-0959.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, IUNLPBEA 2010*, pages 1–9.

Alina Maria Ciobanu and Liviu Dinu. 2014. A Quantitative Insight into the Impact of Translation on Readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR 2014*, pages 104–113.

Meri Coleman and T. L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284.

[5] Repeating the classification experiment for each source language (that is, considering translations from each source language L_i, except for English, one at a time) shows that the differences in performance are not statistically significant ($p < 0.05$). Thus, we conclude that readability features cannot discriminate between original texts and translations significantly better for some of the source languages than for the others.

Kevyn Collins-Thompson. 2011. Enriching Information Retrieval with Reading Level Prediction. In *SIGIR 2011 Workshop on Enriching Information Retrieval*.

Edgar Dale and Jeanne Chall. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

Noemie Elhadad and Komal Sutaria. 2007. Mining a Lexicon of Technical Terms and Lay Equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, BioNLP 2007*, pages 49–56.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING 2010*, pages 276–284.

Thomas François and Eleni Miltsakaki. 2012. Do NLP and Machine Learning Improve Traditional Readability Formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR 2012*, pages 49–57.

Valerio Franchina and Roberto Vacca. 1986. Adaptation of Flesch Readability Index on a Bilingual Text Written by the Same Author both in Italian and English Languages. *Linguaggi*, 3:47–49.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL 2007*, pages 460–467.

Fernandez Huerta. 1959. Medida Sencillas de Lecturabilidad. *Consigna*, 214:29–32.

Lilian Kandel and Abraham Moles. 1958. Application de l'Indice de Flesch a la Langue Française. *Cahiers Etudes de Radio-Television*, 19:253–274.

Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch Report, Millington, TN: Chief of Naval Training.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004*, pages 388–395.

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

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL 2011*, pages 1318–1326.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Harry McLaughlin. 1969. SMOG Grading: a New Readability Formula. *Journal of Reading*, 12(8):639–646.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Christina Schäffner and Susan Bassnett. 2010. Politics, Media and Translation - Exploring Synergies. In *Political Discourse, Media and Translation*, pages 1–29. Newcastle upon Tyne: Cambridge Scholars Publishing.

David J. Sheskin. 2003. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press.

Edgar A. Smith and R. J. Senter. 1967. Automated Readability Index. *Wright-Patterson Air Force Base. AMRL-TR-6620*.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical Machine Translation with Readability Constraints. In *Proceedings of the 19th Nordic Conference on Computational Linguistics, NODALIDA 2013*, pages 375–386.

Yifeng Sun. 2012. Translation and Strategies for Cross-Cultural Communication. *Chinese Translators Journal*, 33(1):16–23.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218.

Anna Trosborg, editor. 1997. *Text Typology and Translation*. Benjamins Translation Library.

Hans van Halteren. 2008. Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics, COLING 2008*, pages 937–944.