Translationese and Its Dialects

Moshe Koppel
Department of Computer Science
Bar Ilan University
Ramat-Gan, Israel 52900
moishk@gmail.com

Noam Ordan
Department of Computer Science
University of Haifa
Haifa, Israel 31905
noam.ordan@gmail.com

Abstract

While it is has often been observed that the product of translation is somehow different than non-translated text, scholars have emphasized two distinct bases for such differences. Some have noted interference from the source language spilling over into translation in a source-language-specific way, while others have noted general effects of the process of translation that are independent of source language. Using a series of text categorization experiments, we show that both these effects exist and that, moreover, there is a continuum between them. There are many effects of translation that are consistent among texts translated from a given source language, some of which are consistent even among texts translated from families of source languages. Significantly, we find that even for widely unrelated source languages and multiple genres, differences between translated texts and non-translated texts are sufficient for a learned classifier to accurately determine if a given text is translated or original.

1 Introduction

The products of translation (written or oral) are generally assumed to be ontologically different from non-translated texts. Researchers have emphasized two aspects of this difference. Some (Baker 1993) have emphasized general effects of the process of translation that are independent of source language and regard the collective product of this process in a given target language as an 'interlanguage' (Selinker, 1972), 'third code' (Frawley, 1984) or 'translationese' (Gellerstam, 1986). Others (Toury, 1995) have emphasized the effects of *interference*, the process by which a specific source language leaves distinct marks or fingerprints in the target language, so that translations from different source languages into the same target language may be regarded as distinct dialects of translationese.

We wish to use text categorization methods to set both of these claims on a firm empirical foundation. We will begin by bringing evidence for two claims:

(1) Translations from different source languages into the same target language are sufficiently different from each other for a learned classifier to accurately identify the source language of a given translated text;

(2) Translations from a mix of source languages are sufficiently distinct from texts originally written in the target language for a learned classifier to accurately determine if a given text is translated or original.

Each of these claims has been made before, but our results will strengthen them in a number of ways. Furthermore, we will show that the degree of difference between translations from two source languages reflects the degree of difference between the source languages themselves. Translations from cognate languages differ from non-translated texts in similar ways, while translations from unrelated languages differ from non-translated texts in distinct ways. The same result holds for families of languages.

The outline of the paper is as follows. In the following section, we show that translations from different source languages can be distinguished from each other and that closely related source languages manifest similar forms of interference. In section 3, we show that, in a corpus involving five European languages, we can distinguish translationese from non-translated text and we consider some salient markers of translationese. In section 4, we consider the extent to which markers of translationese cross over into non-European languages as well as into different genres. Finally, we consider possible applications and implications for future studies.

2 Interference Effects in Translationese

In this section, we perform several text categorization experiments designed to show the extent to which interference affects (both positively and negatively) our ability to classify documents.

2.1 The Europarl Corpus

The main corpus we will use throughout this paper is Europarl (Koehn, 2005), which consists of transcripts of addresses given in the European Parliament. The full corpus consists of texts translated into English from 11 different languages (and vice versa), as well as texts originally produced in English. For our purposes, it will be sufficient to use translations from five languages (Finnish, French, German, Italian and Spanish), as well as original English. We note that this corpus constitutes a *comparable corpus* (Laviosa, 1997), since it contains (1) texts written originally in a certain language (English), as well as (2) texts translated into that same language, matched for genre, domain, publication timeframe, etc. Each of the five translated components is a text file containing just under 500,000 words; the original English component is a file of the same size as the aggregate of the other five.

The five source languages we use were selected by first eliminating several source languages for which the available text was limited and then choosing from among the remaining languages, those of varying degrees of pairwise similarity. Thus, we select three cognate (Romance) languages (French, Italian and Spanish), a fourth less related language (German), and a fifth even further removed (Finnish). As will become clear, the motivation is to see whether the distance between the languages impacts the distinctiveness of the translation product.

We divide each of the translated corpora into 250 equal chunks, paying no attention to natural units within the corpus. Similarly, we divide the original English corpus into 1250 equal chunks. We set aside 50 chunks from each of the translated corpora and 250 chunks from the original English corpus for development purposes (as will be explained below). The experiments described below use the remaining 1000 translated chunks and 1000 original English chunks.

2.2 Identifying source language

Our objective in this section is to measure the extent to which translations are affected by source language. Our first experiment will be to use text categorization methods to learn a classifier that categorizes translations according to source language. We will check the accuracy of such classifiers on out-of-sample texts. High accuracy would reflect that there are exploitable differences among translations of otherwise comparable texts that differ only in terms of source language.

The details of the experiment are as follows. We use the 200 chunks from each translated corpus, as described above. We use as our feature set a list of 300 function words taken from LIWC (Pennebaker, 2001) and represent each chunk as a vector of size 300 in which each entry represents the frequency of the corresponding feature in the chunk. The restriction to function words is crucial; we wish to rely only on stylistic differences rather than content differences that might be artifacts of the corpus.

We use Bayesian logistic regression (Madigan, 2005) as our learning method in order to learn a classifier that classifies a given text into one of five classes representing the different source languages. We use 10-fold cross-validation as our testing method.

We find that 92.7% of documents are correctly classified.

In Table 1 we show the confusion matrix for the five languages. As can be seen, there are more mistakes across the three cognate languages than between those three languages and German and still fewer mistakes involving the more distant Finnish language.

	It	Fr	Es	De	Fi
It	169	19	8	4	0
Fr	18	161	12	8	1
Es	3	11	172	11	3
De	4	12	3	178	3
Fi	0	1	2	5	192

Table 1: Confusion matrix for 10-fold cross validation experiment to determine source language of texts translated into English

This result strengthens that of van Halteren (2008) in a similar experiment. Van Halteren, also using Europarl (but with Dutch as the fifth source language, rather than Finnish), obtained accuracy of 87.2%-96.7% for a two-way decision on source language, and 81.5%-87.4% for a six-way decision (including the original which has no source language). Significantly, though, van Halteren's feature set included content words and he notes that many of the most salient differences reflected differences in thematic emphasis. By restricting our feature set to function words, we neutralize such effects.

In Table 2, we show the two words most over-represented and the two words most under-represented in translations from each source language (ranked according to an unpaired T-test). For each of these, the difference between frequency of use in the indicated language and frequency of use in the other languages in aggregate is significant at p<0.01.

	over-represented	**under-represented**
Fr	*of, finally*	*here, also*
It	*upon, moreover*	*also, here*
Es	*with, therefore*	*too, then*
De	*here, then*	*of, moreover*
Fi	*be, example*	*me, which*

Table 2: Most salient markers of translations from each source language.

The two most underrepresented words for French and Italian, respectively, are in fact identical. Furthermore, the word *too* which is underrepresented for Spanish is a near synonym of *also* which appears in both French and Spanish. This suggests the possibility that interference effects in cognate languages such as French, Italian and Spanish might be similar. We will see presently that this is in fact the case.

When a less related language is involved we see the opposite picture. For German, both underrepresented items appear as overrepresented in the Romance languages, and, conversely, underrepresented items in the Romance languages appear as overrepresented items for German. This may cast doubt on the idea that all translations share universal properties and that at best we may claim that particular properties are shared by closely related languages but not others. In the experiments presented in the next subsection, we'll find that translationese is gradable: closely related languages share more features, yet even further removed languages share enough properties to hold the general translationese hypothesis as valid.

2.3 Identifying translationese per source language

We now wish to measure in a subtler manner the extent to which interference affects translation. In this experiment, the challenge is to learn a classifier that classifies a text as belonging to one of only two classes: original English (O) or translated-into-English (T). The catch is that all our training texts for the class T will be translations from some fixed source language, while all our test documents in T will be translations from a different source language. What accuracy can be achieved in such an experiment? The answer to this question will tell us a great deal about how much of translationese is general and how much of it is language dependent. If accuracy is close to 100%, translationese is purely general (Baker, 1993). (We already know from the previous experiment that that's not the case.). If accuracy is near 50%, there are no general effects, just language-dependent ones. Note that, whereas in our first experiment above pair-specific interference facilitated good classification, in this experiment pair-specific interference is an impediment to good classification.

The details of the experiment are as follows. We create, for example, a "French" corpus consisting of the 200 chunks of text translated from French and 200 original English texts. We similarly create a corpus for each of the other source languages, taking care that each of the 1000 original English texts appears in exactly one of the corpora. As above, we represent each chunk in terms of frequencies of function words. Now, using Bayesian logistic regression, we learn a classifier that distinguishes T from O in the French corpus. We then apply this learned classifier to the texts in, for example, the equivalent "Italian" corpus to see if we can classify them as translated or original. We repeat this for each of the 25 ⟨train_corpus, test_corpus⟩ pairs.

In Table 3, we show the accuracy obtained for each such pair. (For the case where the training corpus and testing corpus are identical – the diagonal of the matrix – we show results for ten-fold cross-validation.)

We note several interesting facts. First, results of cross-validation within each corpus are very strong. For any given source language, it is quite easy to distinguish translations from original English. This corroborates results obtained by Baroni and Bernardini (2006), Ilisei et al. (2010), Kurokawa et al. (2009) and van Halteren (2008), which we will discuss below.

We note further, that for the cases where we train on one source language and test on another, results are far worse. This clearly indicates that interference effects from one source language might be misleading when used to identify translations from a different language. Thus, for example, in the Finnish corpus, the word *me* is a strong indicator of original English (constituting 0.0003 of tokens in texts translated from Finnish

Train

	It	Fr	Es	De	Fi
It	98.3	91.5	86.5	71.3	61.5
Fr	91	97	86.5	68.5	60.8
Es	84.5	88.3	95.8	76.3	59.5
De	82	83.3	78.5	95	80.8
Fi	56	60.3	56	62.3	97.3

Table 3: Results of learning a T vs. O classifier using one source language and testing it using another source language

as opposed to 0.0015 of tokens in original English texts), but in the German corpus, *me* is an indicator of translated text (constituting 0.0020 of tokens in text translated from German).

The most interesting result that can be seen in this table is that the accuracy obtained when training using language x and testing using language y depends precisely on the degree of similarity between x and y. Thus, for training and testing within the three cognate languages, results are fairly strong, ranging between 84.5% and 91.5%. For training/testing on German and testing/training on one of the other European languages, results are worse, ranging from 68.5% to 83.3%. Finally, for training/testing on Finnish and testing/training on any of the European languages, results are still worse, hovering near 60% (with the single unexplained outlier for training on German and testing on Finnish).

Finally, we note that even in the case of training or testing on Finnish, results are considerably better than random, suggesting that despite the confounding effects of interference, some general properties of translationese are being picked up in each case. We explore these in the following section.

3 General Properties of Translationese

Having established that there are source-language-dependent effects on translations, let’s now consider source-language-independent effects on translation.

3.1 Identifying translationese

In order to identify general effects on translation, we now consider the same two-class classification problem as above, distinguishing T from O, except that now the translated texts in both our train and test data will be drawn from multiple source languages. If we succeed at this task, it must be because of features of translationese that cross source-languages.

The details of our experiment are as follows. We use as our translated corpus, the 1000 translated chunks (200 from each of five source languages) and as our original English corpus all 1000 original English chunks. As above, we represent each chunk in terms of function words frequencies. We use Bayesian logistic regression to learn a two-class classifier and test its accuracy using ten-fold cross-validation.

Remarkably, we obtain accuracy of 96.7%.

This result extends and strengthens results reported in some earlier studies. Ilisei et al. (2010), Kurokawa (2009) and van Halteren (2008) each obtained above 90% accuracy in distinguishing translation from original. However, in each case the translations were from a single source language. (Van Halteren considered multiple source languages, but each learned classifier used only one of them.) Thus, those results do not prove that translationese has distinctive source-language-independent features. To our knowledge, the only earlier work that used a learned classifier to identify translations in which both test and train sets involved multiple source languages is Baroni and Bernardini (2006), in which the target language was Italian and the source languages were known to be varied. The actual distribution of source languages was, however, not known to the researchers. They obtained accuracy of 86.7%. Their result was obtained using combinations of lexical and syntactic features.

3.2 Some distinguishing features

Let us now consider some of the most salient function words for which frequency of usage in T differs significantly from that in O. While there are many such features, we focus on two categories of words that are most prominent among those with the most significant differences.

First, we consider animate pronouns. In Table 4, we show the frequencies of animate pronouns in O and T, respectively (the possessive pronouns, *mine*, *yours* and *hers*, not shown, are extremely rare in the corpus). As can be seen, all pronouns are under-represented in T; for most (bolded), the difference is significant at $p<0.01$.

By contrast, the word *the* is significantly overrepresented in T (15.32% in T vs. 13.73% in O; significant at $p<0.01$).

word	freq O	freq T
I	**2.552%**	2.148%
we	**2.713%**	2.344%
you	0.479%	0.470%
he	**0.286%**	0.115%
she	**0.081%**	0.039%
me	0.148%	0.141%
us	**0.415%**	0.320%
him	**0.066%**	0.033%
her	**0.091%**	0.056%
my	**0.462%**	0.345%
our	**0.696%**	0.632%
your	0.119%	0.109%
his	**0.218%**	0.123%

Table 4: Frequency of pronouns in O and T in the Europarl corpus. Bold indicates significance at p<0.01.

In Table 5, we consider cohesive markers, tagged as adverbs (Schmid, 2004). (These are adverbs that can appear at the beginning of a sentence followed immediately by a comma.)

word	freq O	freq T
therefore	0.153%	**0.287%**
thus	0.015%	**0.041%**
consequently	0.006%	**0.014%**
hence	0.007%	**0.013%**
accordingly	0.006%	**0.011%**
however	0.216%	**0.241%**
nevertheless	0.019%	**0.045%**
also	0.460%	**0.657%**
furthermore	0.012%	**0.048%**
moreover	0.008%	**0.036%**
indeed	**0.098%**	0.053%
actually	**0.065%**	0.042%

Table 5: Frequency of cohesive adverbs in O and T in the Europarl corpus. Bold indicates significance at p<0.01.

We note that the preponderance of such cohesive markers are significantly more frequent in translations. In fact, we also find that a variety of phrases that serve the same purpose as cohesive adverbs, such as *in fact* and *as a result* are significantly more frequent in translationese.

The general principle underlying these phenomena is subject to speculation. Previous researchers have noted the phenomenon of *explicitation*, according to which translators tend to render implicit utterances in the source text into explicit utterances in the target text (Blum-Kulka, 1986, Laviosa-Braithwaite, 1998), for example by filling out elliptical expressions or adding connectives to increase cohesion of the text (Laviosa-Braithwaite, 1998). It is plausible that the use of cohesive adverbs is an instantiation of this phenomenon.

With regard to the under-representation of pronouns and the over-representation of *the*, there are a number of possible interpretations. It may be that this too is the result of explicitation, in which anaphora is resolved by replacing pronouns with noun phrases (e.g., *the man* instead of *he*). But it also might be that this is an example of *simplification* (Laviosa- Braithwaite 1998, Laviosa 2002), according to which the translator simplifies the message, the language, or both. Related results confirming the simplification hypothesis were found by Ilisei et al. (2010) on Spanish texts. In particular, they found that type-to-token ratio (*lexical variety*/*richness*), mean sentence length and proportion of grammatical words (*lexical density*/*readability*) are all smaller in translated texts.

We note that Van Halteren (2008) and Kurokawa et al. (2009), who considered lexical features, found cultural differences, like over-representation of *ladies* and *gentlemen* in translated speeches. Such differences, while of general interest, are orthogonal to our purposes in this paper.

3.3 Overriding language-specific effects

We found in Section 2.3 that when we trained in one language and tested in another, classification succeeded to the extent that the source languages used in training and testing, respectively, are related to each other. In effect, general differences between translationese and original English were partially overwhelmed by language-specific differences that held for the training language but not the test language. We thus now revisit that earlier experiment, but restrict ourselves to features that distinguish translationese from original English generally.

To do this, we use the small development corpus described in Section 2.1. We use Bayesian logistic regression to learn a classifier to distinguish between translationese and original English. We select the 10 highest-weighted function-word markers for T and the 10 highest-weighted function-word markers for O in the development corpus. We then rerun our train-on-source-language-x, test-on-source-language-y experiment using this restricted set as our feature set. We now find that even in the difficult case where we train on Finnish and test on another language (or vice versa), we succeed at distinguishing translationese from original English with accuracy above 80%. This considerably improves the earlier results shown in Table 3. Thus, a bit of feature engineering facilitates learning a good classifier for T vs. O even across source languages.

4 Other Genres and Language Families

We have found both general and language-specific differences between translationese and original English in one large corpus. It might be wondered whether the phenomena we have found hold in other genres and for a completely different set of source languages. To test this, we consider a second corpus.

4.1 The IHT corpus

Our second corpus includes three translated corpora, each of which is an on-line local supplement to the International Herald Tribune (IHT): *Kathimerini* (translated from Greek), *Ha'aretz* (translated from Hebrew), and the *JoongAng Daily* (translated from Korean). In addition, the corpus includes original English articles from the IHT. Each of the four components contains four different domains balanced roughly equally: news (80,000 words), arts and leisure (50,000), business and finance (50,000), and opinion (50,000) and each covers the period from April-September 2004. Each component consists of about 230,000 tokens. (Unlike for our Europarl corpus, the amount of English text available is not equal to the aggregate of the translated corpora, but rather equal to each of the individual corpora.)

It should be noted that the IHT corpus belongs to the writing modality while the Europarl corpus belongs to the speaking modality (although possibly post-edited). Furthermore, the source languages (Hebrew, Greek and Korean) in the IHT corpus are more disparate than those in the Europarl corpus.

Our first objective is to confirm that the results we obtained earlier on the Europarl corpus hold for the IHT corpus as well.

Perhaps more interestingly, our second objective is to see if the gradability phenomenon observed earlier (Table 3) generalizes to families of languages. Our first hypothesis is that a classifier for identifying translationese that is trained on Europarl will succeed only weakly to identify translationese in IHT. But our second hypothesis is that there are sufficient general properties of translationese that cross language families and genres that a learned classifier can accurately identify translationese even on a test corpus that includes both corpora, spanning eight disparate languages across two distinct genres.

4.2 Results on IHT corpus

Running essentially the same experiments as described for the Europarl corpus, we obtain the following results.

First of all, we can determine source language with accuracy of 86.5%. This is a somewhat weaker result than the 92.7% result obtained on Europarl, especially considering that there are only three classes instead of five. The difference is most likely due to the fact that the IHT corpus is about half the size of the Europarl corpus. Nevertheless, it is clear that source language strongly affects translationese in this corpus.

Second, as can be seen in Table 6, we find that the gradability phenomenon occurs in this corpus as well. Results are strongest when the train and test corpora involve the same source language and trials involving Korean, the most distant language, are somewhat weaker than those across Greek and Hebrew.

Train

	Gr	He	Ko
Gr	89.8	73.4	64.8
He	82.0	86.3	65.5
Ko	73.0	72.5	85.0

Table 6: Results of learning a T vs. O classifier using one source language and testing it using another source language

Third, we find in ten-fold cross-validation experiments that we can distinguish translationese from original English in the IHT corpus with accuracy of 86.3%. Thus, despite the great distance between the three source languages in this corpus, general differences between translationese and original English are sufficient to facilitate reasonably accurate identification of translationese.

4.3 Combining the corpora

First, we consider whether a classifier learned on the Europarl corpus can be used to identify translationese in the IHT corpus, and vice versa. It would be consistent with our findings in Section 2.3, that we would achieve better than random results but not high accuracy, since there are no doubt features common to translations from the five European languages of Europarl that are distinct from those of translations from the very different languages in IHT.

In fact, we find that training on Europarl and testing on IHT yields accuracy of 64.8%, while training on IHT and testing on Europarl yields accuracy of 58.8%. The weak results reflect both differences between the families of source languages involved in the respective corpora, as well as genre differences. Thus, for example, we find that of the pronouns shown in Table 4 above, only *he* and *his* are significantly under-represented in translationese in the IHT corpus. Thus, that effect is specific either to the genre of Europarl or to the European languages considered there.

Now, we combine the two corpora and check if we can identify translationese across two genres and eight languages. We run the same experiments as described above, using 200 texts from each of the eight source languages and 1600 non-translated English texts, 1000 from Europarl and 600 from IHT.

In 10-fold cross-validation, we find that we can distinguish translationese from non-translated English with accuracy of 90.5%.

This shows that there are features of translationese that cross genres and widely disparate languages. Thus, for one prominent example, we find that, as in Europarl, the word *the* is over-represented in translationese in IHT (15.36% in T vs. 13.31% in O; significant at p<0.01). In fact, the frequencies across corpora are astonishingly consistent.

To further appreciate this point, let’s look at the frequencies of cohesive adverbs in the IHT corpus.

We find essentially, the same pattern in IHT as we did in Europarl. The preponderance of cohesive adverbs are over-represented in translationese, most of them with differences significant at p<0.01. Curiously, the word *actually* is a counter-example in both corpora.

word	freq O	freq T
therefore	0.011%	**0.031%**
thus	0.011%	**0.027%**
consequently	0.000%	**0.004%**
hence	0.003%	0.007%
accordingly	0.003%	0.003%
however	0.078%	**0.129%**
nevertheless	0.008%	**0.018%**
also	0.305%	**0.453%**
furthermore	0.003%	**0.011%**
moreover	0.009%	0.008%
indeed	0.018%	0.024%
actually	0.032%	0.018%

Table 7: Frequency of cohesive adverbs in O and T in the IHT corpus. Bold indicates significance at p<0.01.

5 Conclusions

We have found that we can learn classifiers that determine source language given a translated text, as well as classifiers that distinguish translated text from non-translated text in the source language. These text categorization experiments suggest that both source language and the mere fact of being translated play a crucial role in the makeup of a translated text.

It is important to note that our learned classifiers are based solely on function words, so that, unlike earlier studies, the differences we find are unlikely to include cultural or thematic differences that might be artifacts of corpus construction.

In addition, we find that the exploitability of differences between translated texts and non-translated texts are related to the difference between source languages: translations from similar source languages are different from non-translated texts in similar ways.

Linguists use a variety of methods to quantify the extent of differences and similarities between languages. For example, Fusco (1990) studies translations between Spanish and Italian and considers the impact of structural differences between the two languages on translation quality. Studying the differences and distance between languages by comparing translations into the same language may serve as another way to deepen our typological knowledge. As we have seen, training on source language x and testing on source language y provides us with a good estimation of the distance between languages, in accordance with what we find in standard works on typology (cf. Katzner, 2002).

In addition to its intrinsic interest, the finding that the distance between languages is directly correlated with our ability to distinguish translations from a given source language from non-translated text is of great importance for several computational tasks. First, translations can be studied in order to shed new light on the differences between languages and can bear on attested techniques for using cognates to improve machine translation (Kondrak & Sherif, 2006). Additionally, given the results of our experiments, it stands to reason that using translated texts, especially from related source languages, will prove beneficial for constructing language models and will outperform results obtained from non-translated texts. This, too, bears on the quality of machine translation.

Finally, we find that there are general properties of translationese sufficiently strong that we can identify translationese even in a combined corpus that is comprised of eight very disparate languages across two distinct genres, one spoken and the other written. Prominent among these properties is the word *the*, as well as a number of cohesive adverbs, each of which is significantly over-represented in translated texts.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Gill Francis Mona Baker and Elena Tognini Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233-252. John Benjamins, Amsterdam.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing,* 21(3):259-274.

Shoshan Blum-Kulka. Shifts of cohesion and coherence in translation. 1986. In Juliane House and Shoshana Blum-Kulka (Eds), *Interlingual and Intercultural Communication* (17-35). Tübingen: Günter Narr Verlag.

William Frawley. 1984. Prolegomenon to a theory of translation. In William Frawley (ed), *Translation. Literary, Linguistic and Philosophical Perspectives* (179-175). Newark: University of Delaware Press.

Maria Antonietta Fusco. 1990. Quality in conference interpreting between cognate languages: A preliminary approach to the Spanish-Italian case. *The Interpreters' Newsletter*, 3, 93-97.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English, in Lars Wollin & Hans Lindquist (eds.), *Translation Studies in Scandinavia* (88-95). Lund: CWK Gleerup.

I. I. Ilisei, D. Inkpen, G. Pastor, G., and Mitkov, R. 2010. Identification of Translationese: A Supervised Learning Approach, *CICLing 2010,* pp. 503-511.

Kenneth Katzner. 2002. *The Languages of the World*. Routledge.

Grzegorz Kondrak and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of the Workshop on Linguistic Distances* (LD '06). 43-50.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In Proceedings of MT-Summit XII.

Sara Laviosa: 1997. *How Comparable can 'Comparable Corpora' Be?*. Target, 9 (2), pp. 289-319.

Sara Laviosa-Braithwaite. 1998. In Mona Baker (ed.) *Routledge Encyclopedia of Translation Studies*. London/New York: Routledge, pp.288-291.

Sara Laviosa. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam/New York: Rodopi.

David Madigan, Alexander Genkin, David D. Lewis and Dmitriy Fradkin 2005. Bayesian Multinomial Logistic Regression for Author Identification, In Maxent Conference, 509-516.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Erlbaum Publishers, Mahwah, NJ, USA.

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. 2004. In *Proceedings of International Conference on New Methods in Language Processing*.

Larry Selinker.1972. Interlanguage. *International Review of Applied Linguistics*. 10, 209-241.

Gideon Toury. 1995. *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam / Philadelphia.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937-944.