On the Features of Translationese

Vered Volansky, Noam Ordan, and Shuly Wintner*
Department of Computer Science, University of Haifa
Mount Carmel, 31905 Haifa, Israel

Abstract

Much research in translation studies indicates that translated texts are ontologically different from original, non-translated ones. Translated texts, in any language, can be considered a dialect of that language, known as 'translationese'. Several characteristics of translationese have been proposed as universal in a series of hypotheses. In this work we test these hypotheses using a computational methodology that is based on supervised machine learning. We define several classifiers that implement various linguistically-informed features, and assess the degree to which different sets of features can distinguish between translated and original texts. We demonstrate that some feature sets are indeed good indicators of translationese, thereby corroborating some hypotheses, whereas others perform much worse (sometimes at chance level), indicating that some 'universal' assumptions have to be reconsidered.

1 Introduction

This work addresses the differences between translated (T) and original (O), non-translated texts. These differences, to which we refer as 'features', have been discussed and studied extensively by translation scholars in the last three decades. In this work we employ computational means to investigate them quantitatively. Focusing only on English, our main methodology is based on *machine learning*, more specifically an application of machine learning to *text classification*.

The special status of translated texts is a compromise between two forces, fidelity to the source text, on the one hand, and fluency in the target language, on the other hand. Both forces exist simultaneously: some 'fingerprints' of the source text are left on the target text, and at the same time the translated text includes shifts from the source so as to be more fluent and produce a better fit to the target language model. The differences between O and T were studied empirically since the 1990s by translation scholars on

*This is the authors' pre-print copy which differs from the final publication.

computerized corpora (Laviosa, 2002), but only recently, since Baroni and Bernardini (2006), has it been shown that distinguishing between O and T can be done automatically with a high level of accuracy.

Toury (1980) paved the way for studying translated texts in comparison to *target language* texts, ignoring the source text altogether. The idea behind this move was that translations as such, regardless of the source language, have something in common, certain stylistic features governed by so-called translation *norms*, and therefore, in order to learn about these special marks of translation, the right point of reference is non-translated texts.

Baker (1993) calls for compiling and digitizing 'comparable corpora' and using them to study 'translation universals', such as *simplification*, the tendency to make the source text simpler lexically, syntactically, etc, or *explicitation*, the tendency to render implicit utterances in the original more explicit in the translation. This call sparked a long-lasting quest for translation universals, and several works test such hypotheses in many target languages, including English, Finnish, Hungarian, Italian and Swedish (Mauranen and Kujamäki, 2004; Mauranen, 2008).

In this study we refrain from the token 'universal' and focus instead on 'features'. This terminological choice has several reasons. First, the focus is mostly on data and empirical findings, and less on translation theory as such. Whereas the features are motivated by and organized according to theoretical categories, we admit that certain features can belong to more than one theoretical category (see Section 6). Second, we show that certain features (such as mean sentence length) are highly dependent on the source language, and in general many of the features have a skewed distribution (again, see Section 6); we therefore cast doubt on the universality of 'universals'. Third, the term 'feature' (or sometimes 'attribute') is common in machine learning parlance, which is the main methodology used in this study.

This paper uses machine learning algorithms to distinguish between O and T. In particular, we apply *text classification*, a methodology that has been used for classifying texts according to topic, genre, etc. (Sebastiani, 2002), but also for *authorship attribution*, classification of texts according to their authors' gender, age, provenance, and more (Koppel et al., 2009). This methodology has been successfully applied to studying O vs. T in various datasets and in different source and target languages (see Section 2). In most of these works the focus is on computational challenges, namely classifying with high accuracy, expanding to more scenarios (for example, cross-domain classification), and minimizing the samples on which the computer is trained, so the attribution can be done on smaller portions of texts. Our study, in contrast, employs this methodology to examine a list of 32 features of 'translationese' suggested by translation scholars, with an eye to the question whether some of these features can be utilized to tell O from T.

The main contribution of this work is thus theoretical, rather than prac-

tical: we use computational means to investigate several translation studies hypotheses, corroborating some of them but refuting others. More generally, we advocate the use of automatic text classification as a methodology for investigating translation studies hypotheses in general, and translation universals in particular.

After reviewing related work in the next section, we detail our methodology in Section 3. We describe several translation studies hypotheses in Section 4, explaining how we model them computationally in terms of features used for classification. The results of the classifiers are reported in Section 5, and are analyzed and discussed in Section 6. We conclude with suggestions for future research.

2 Related Work

Numerous studies suggest that translated texts differ from original ones. Gellerstam (1986) compares texts written originally in Swedish and texts translated from English into Swedish. He notes that the differences between them do not indicate poor translation but rather a statistical phenomenon, which he terms *translationese*. The features of translationese were theoretically organized under the terms *laws of translation* or *translation universals*.

Toury (1980, 1995) distinguishes between two laws: the *law of interference* and the *law of growing standardization*. The former pertains to the fingerprints of the source text that are left in the translation product. The latter pertains to the effort to standardize the translation product according to existing norms in the target language and culture. The combined effect of these laws creates a hybrid text that partly corresponds to the source text and partly to texts written originally in the target language, but in fact is neither of them (Frawley, 1984).

Baker (1993) suggests several candidates for translation universals, which are claimed to appear in any translated text, regardless of the source language: "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1993, p. 243). Consequently, there is no need to study translations vis-à-vis their source. The corpus needed for such study is termed *comparable corpus*,[1] where translations from various source languages are studied in comparison to non-translated texts in the same language, holding for genre, domain, time frame, etc. Among the better known universals are *simplification* and *explicitation*, defined and discussed thoroughly by Blum-Kulka and Levenston (1978, 1983) and Blum-Kulka (1986), respectively.

Following Baker (1993), a quest for the holy grail of translation universals

[1]This term should be distinguished from *comparable corpus* in computational linguistics, where it refers to texts written in *different* languages that contain similar information.

began, culminating in Mauranen and Kujamäki (2004). Chesterman (2004) distinguishes between S-universals and T-universals. S-universals are features that can be traced back to the source text, and include, among others, *lengthening*, *interference*, *dialect normalization* and *reduction of repetitions*. T-universals, on the other hand, are features that should be studied vis-à-vis non-translated texts in the target language, i.e., by using a comparable corpus. These include such features as *simplification*, *untypical patterning* and *under-representation of target-language-specific items*. This distinction classifies putative translation universals into two categories, each of which calls for a different kind of corpus, parallel for S-universals and comparable for T-universals. We cast all our features in a comparable corpus setting (T-universals). Our assumption is that if a feature is reflected in translations from several languages into English, it is very likely present in the source texts from which these translations were generated. Future study could very well verify this assumption.

In the last decade, corpora have been used extensively to study translationese. For example, Al-Shabab (1996) shows that translated texts exhibit lower lexical variety (type-to-token ratio) than originals; Laviosa (1998) shows that their mean sentence length is lower, as is their lexical density (ratio of content to non-content words). Both these studies provide evidence for the simplification hypothesis. Corpus-based translation studies became a very prolific area of research (Laviosa, 2002).

Text classification methods have only recently been applied to the task of identifying translationese. Baroni and Bernardini (2006) use a two-million-token Italian newspaper corpus, in which 30% of the texts are translated from various source languages the proportions of which are not reported. They train a *support vector machine* (SVM) classifier using unigrams, bigrams and trigrams of surface forms, lemmas and part-of-speech (POS) tags. They also experiment with a *mixed* mode, in which function words are left intact but content words are replaced by their POS tags. The best accuracy, 86.7%, is obtained using a combination of lemma and mixed unigrams, bigrams and POS trigrams. Extracting theoretically interesting features, they show that Italian T includes more 'strong' pronouns, implying that translating from non-pro-drop languages to a pro-drop one, like Italian, is marked on T. In other words, if a certain linguistic feature is mandatory in the source language and optional in the target language, more often than not it will be carried over to the target text.

This is a clear case of *positive interference*, where features that do exist in O have greater likelihood to be selected in T. In contrast, there are cases of *negative interference*, where features common in O are under-represented in T (Toury, 1995), and more generally, "[t]ranslations tend to under-represent target-language-specific, unique linguistic features and over-represent features that have straightforward translation equivalents which are frequently used in the source language" (Eskola, 2004, p. 96). Note that Baroni and

Bernardini (2006) use lemmas as features; this can artificially inflate the accuracy of the classifier since lemmas reflect topic and domain information rather than structural differences between the two classes of texts.

Inspired by Baroni and Bernardini (2006), Kurokawa et al. (2009) use a mixed text representation in which content words are replaced by their corresponding POS tags, while function words are retained. The corpus here is the Canadian Hansard, which consists of texts in English and Canadian French and translations in both directions, drawn from official records of the proceedings of the Canadian Parliament. Classification is performed at both the document and the sentence level. Interestingly, they demonstrate that learning the direction is relevant for statistical machine translation: they train systems to translate between French and English (and vice versa) using a French-translated-to-English parallel corpus, and then an English-translated-to-French one. They find that in translating into French it is better to use the latter parallel corpus, and when translating into English it is better to use the former. The contribution of knowledge of the translation direction to machine translation is further corroborated in a series of works (Lembersky et al., 2011, 2012a,b).

van Halteren (2008) shows that there are significant differences between texts translated from different source languages to the same target language in EUROPARL (Koehn, 2005). The features are 1–3-grams of tokens that appear in at least 10% of the texts of each class. There are 6×6 classes: an original and 5 translations from and into the following: Danish, English, French, German, Italian and Spanish. Tokens appearing in less than 10% of the texts in each class are replaced with ⟨x⟩. Thus, for example, *are_right_*⟨x⟩ is a marker of translations form German, while *conditions_of_*⟨x⟩ is a marker of translations from French. The 10% threshold does not totally exclude content words, and therefore many markers reflect cultural differences, most notably the form of address *ladies and gentlemen* which is highly frequent in the translations but rare in original English.

Ilisei et al. (2010) test the simplification hypothesis using machine learning algorithms. As noted earlier, certain features, such as average sentence length, do not provide a rich model, and cannot, by themselves, discriminate between O and T with high accuracy. Therefore, in addition to the 'simplification features', the classifier is trained on POS unigrams, and then each simplification feature is included and excluded and the success rate in both scenarios is compared. They then conduct a *t*-test to check whether the difference is statistically significant.

Ilisei et al. (2010) define several 'simplification features', including average sentence length; sentence depth (as depth of the parse tree); ambiguity (the average number of senses per word); word length (the proportion of syllables per word); lexical richness (type/token ratio); and information load (the proportion of content words to tokens). Working on Spanish, the most informative feature for the task is lexical richness, followed by sentence

length and the proportion of function words to content words. Both lexical richness and sentence length are among the simplification features and are therefore considered to be indicative of the simplification hypothesis. All in all, they succeed in differentiating between translated and non-translated texts with 97.62% accuracy and conclude that simplification features exist and heavily influence the results. These results pertain to Spanish translated from English; Ilisei and Inkpen (2011) extend the results to Romanian, using by and large the same methodology, albeit with somewhat more refined features. Furthermore, Ilisei (2013) experiments also with the explicitation hypothesis (in Spanish and Romanian), defining mainly features whose values are the proportion of some part of speech categories in the text.

Our work is similar in methodology, but is much broader in scope. While Ilisei et al. (2010); Ilisei and Inkpen (2011) use their simplification features to boost the accuracy of the classifier, our goal is different, as we are interested not in the actual accuracy of any feature by itself, but in its contribution, if any, to the classification and translation process. We test some of the simplification features on English translated from *ten* source languages vs. original English. We also add more simplification features to those introduced by Ilisei et al. (2010); Ilisei and Inkpen (2011) to test the simplification hypothesis. Most importantly, we add many more features that test a large array of other hypotheses.

Koppel and Ordan (2011) aim to identify the source language of texts translated to English from several languages, and reason about the similarities or differences of the source languages with respect to the accuracy obtained from this experiment. The data are taken from the EUROPARL corpus, and include original English texts as well as English texts translated from Finnish, French, German, Italian and Spanish. In order to abstract from content, the only features used for classification are frequencies of function words. Koppel and Ordan (2011) can distinguish between original and translated texts with 96.7% accuracy; they can identify the original language with 92.7% accuracy; and they can train a classifier to distinguish between original English and English translated from language L_1, and then use the same classifier to differentiate between original English and English translated from L_2, with accuracies ranging from 56% to 88.3%. Interestingly, the success rate improves when L_1 and L_2 are typologically closer. Thus, training on one Romance language and testing on another yields excellent results between 84.5%-91.5% (there are 6 such pairs for French, Italian and Spanish).

The poor results (56%) of training on T from Finnish and testing on T from Italian or Spanish, for example, cast doubt on the concept of 'translation universals'. It shows that translationese is highly dependent on the pair of languages under study. Although Koppel and Ordan (2011) manage to train on all the T components vs. O and achieve a good result distinguishing between O and T (92.7%), it is exactly their main finding of pair-specific

dependence that may tie this success to their corpus design: three fifths of their corpus belong to the same language family (Romance), another fifth of translations from German is also related (Germanic), and only the last fifth, Finnish, is far removed (Finno-Ugric). In our experiments we use a wider range of source languages in an effort to neutralize this limitation: Romance (Italian, Portuguese, Spanish, and French); Germanic (German, Danish, and Dutch); Hellenic (Greek); and Finno-Ugric (Finnish).

Popescu (2011), too, identifies translationese with machine-learning methods. He uses a corpus from the literary domain, mainly books from the nineteenth century. The corpus contains 214 books, half of which (108) are originally written in British and American English. The other half is of translated English, 76 from French and 30 from German. The book domains are varied and translations are ensured to be of at least minimal quality. Popescu (2011) uses character sequences of length 5, ignoring sentence boundaries, for classification. He achieves 99.53% to 100% accuracy using different cross-validation methods. When training on British English and translations from French, and testing on American English and translations from German, the accuracy is 45.83%. He then uses the original French corpus to eliminate proper names, still at the character level, and achieves 77.08% accuracy. By mixing American and British texts, 76.88% accuracy is achieved.

This work has many advantages from the engineering point of view: extracting characters is a trivial text-processing task; the methodology is language-independent, and with some modifications it can be applied, for example, to Chinese script, where segmenting words is a non-trivial task; it *does not* impose any theoretical notions on the classification; last, the model for O and T is very rich since there are many possible character n-gram values (like *the*, *of*, *-ion*, -ly, etc.) and therefore the model can fit different textual scenarios on which it is tested. Similarly to Popescu (2011), we use simple n-gram characters, $n = 1, 2, 3$, among many other features. The higher n is, the more we can learn about translationese, as we show in Section 4. Still, it should be noted that character n-grams can capture lexical information, which, like lemmas, may reflect topic and domain information rather than structure.

In contrast to some previous works, we use the machine-learning methodology with great care. First, we compile a corpus with multiple source languages, from diverse language families; we balance the proportion of each language within the corpus, and provide detailed information that can be used for replicating our results. Second, we totally abstract away from content so as to be unbiased by the topics of the corpora. A classifier that uses as features the words in the text, for example, is likely to do a good job telling O from T simply because certain words are culturally related to the source language from which the texts are translated (e.g., the word *Paris* in texts translated from French). We provide data on such classifiers, but only

as a "sanity check". Furthermore, while previous works used this methodology to investigate the simplification hypothesis, we use it to investigate a wide array of translation studies hypotheses, including simplification, explicitation, normalization and interference. Finally, and most importantly, we use a plethora of linguistically informed features to learn more about the nature of translationese.

3 Methodology

Our main goal in this work is to study the features of translated texts.[2] Our methodology is corpus-based, but instead of computing quantitative measures of O and T directly, we opt for a more sophisticated, yet more revealing methodology, namely training classifiers on various features, and investigating the ability of different features to accurately distinguish between O and T. We now detail the methodology and motivate it.

3.1 Text Classification with Machine Learning

In *supervised machine-learning*, a *classifier* is trained on labeled examples the classification of which is known a priori. The current task is a binary one, namely there are only two classes: O and T. Each instance in the two classes has to be *represented*: a set of numeric *features* is extracted from the instances, and a generic machine-learning algorithm is then trained to distinguish between *feature vectors* representative of one class and those representative of the other. For example, one set of features for natural texts could be the words (or tokens) in the text; the *values* of these features are the number of occurrences of each word in the instance. This set of features is extracted from the text instances in both classes, and then each of the classes is modeled differently such that there is a model for how O should look like and a model for how T should look like. Given enough data for training and given that the features are indeed relevant, the trained classifier can then be given an 'unseen' text, namely a text that is not included in the training set. Such a text is again represented by a feature vector in the same manner, and the classifier can predict whether it belongs to the O class or to the T class. Such unseen texts are known as "test set".

One important property of such classifiers is that they assign "weights" to the features used for classification, such that significant features are assigned higher weights. Due to potential dependencies among features, some features may be assigned weights that diminish their importance on their own, as they do not add any important data to the classifier. This means

[2] A terminological note is in place: throughout this paper, O and T refer to texts written in the same language, specifically in English. The languages from which T was translated are therefore referred to as the *source* languages. When we say *French*, for example, we mean texts translated to English from French.

that low weights are not always very reliable; but if a feature is assigned a high weight, it is certainly a good indication of a significant difference between the two classes (the inverse does not necessarily hold).

3.2 Motivation

Applying machine learning algorithms to identify the class of the text (O or T) is thus a sound methodology for assessing the predictive power of a feature set. This is by no means a call to abandon traditional significance tests as a tool to learn about differences between texts, and in fact, we use both in this study. But text classification is more robust, in the sense that it reflects not just average differences between classes, but also the different distributions of features across the classes, in a way that facilitates *generalization*: prediction of the class of new, unseen examples.

To illustrate this point, consider the case of punctuation marks. We compare the frequencies of several marks in O and T. Table 1 summarizes the data: for each punctuation mark, it lists the relative frequency (per 1000 token) in O and T; the ratio between O and T ('ratio'); whether the feature in question typifies O or T according to a log-likelihood (LL) test ($p < 0.05$); and the strength of the weight assigned to the feature by a particular classifier (Section 3.3), where T1 is the most prominent feature of translation (the one with the highest weight, as determined by the classifier), T2 the second most prominent, and so on; the same notation from O1 to O7 is applied to O.

	Frequency				
Mark	**O**	**T**	**Ratio**	**LL**	**Weight**
,	37.83	49.79	0.76	T	T1
(	0.42	0.72	0.58	T	T2
’	1.94	2.53	0.77	T	T3
)	0.40	0.72	0.56	T	T4
/	0.30	0.30	1.00	—	—
[	0.01	0.02	0.45	T	—
]	0.01	0.02	0.44	T	—
”	0.33	0.22	1.46	O	O7
!	0.22	0.17	1.25	O	O6
.	38.20	34.60	1.10	O	O5
:	1.20	1.17	1.03	—	O4
;	0.84	0.83	1.01	—	O3
?	1.57	1.11	1.41	O	O2
-	2.68	2.25	1.19	O	O1

Table 1: Summary data for punctuation marks across O and T

The most prominent marker of T according to the classifier is the comma, which is indeed about 1.3 times more frequent in T than in O. There are punctuation marks for which the ratio is much higher; for example, square brackets are about 2.2 times more frequent in T. But their frequency in the corpus is very low, and therefore, this difference is not robust enough to make a prediction. Theoretically it may be interesting to note that in translations from Swedish into English, for example, there are four times more square brackets than in original English, but this plays no significant role in the classification task.

Conversely, there are cases where the critical value of LL is not significant by common standards, but it does play a role in classification. Such is the case of the colon. The ratio O/T is 1.03 and the critical value is 2.06, namely $p < 0.15$. This value still accounts for 85% of the cases and although common statistical wisdom would rule it out as an insignificant feature, it does play a significant role in telling O from T using text classification techniques.

The parentheses appear almost always together. We notice rare cases of '(' appearing in itself, usually as a result of tokenization problem, and some cases of the right parenthesis ')' appearing by itself, usually in enumerating items within a paragraph, a common notation in non-English languages and therefore three times more frequent in T than in O (30 vs. 10 cases, respectively). Although the raw frequency of both '(' and ')' is about the same and although the ratio between their frequency in O and their frequency in T is nearly identical, '(' appears to be a better marker of T according to the classifier. When a classifier is encountered with two highly dependent features it may ignore one of them altogether. This does not mean the ignored feature is not important, it only means it does not add much new information.

In summary, we use text classification algorithms to measure the robustness of each feature set. We are interested in differences between O and T, but we are also interested in finding out how revealing these features are, how prominent in the marking of translated text, to the effect that they have a predictive power. We use the information produced by the classifiers to provide a preliminary analysis. Then, to make a finer analysis, we check some of the features manually and conduct significance tests. We believe that using text classification techniques provides a good tool to study the makeup of translated texts in a general way, on the one hand, and that using statistical significance tests on occasion enables the researcher to look at less frequent events which are no doubt part of the story of translationese, on the other hand.

3.3 Experimental Setup

The main corpus we use is the proceedings of the European Parliament, EUROPARL (Koehn, 2005), with approximately 4 million tokens in English

(O) and the same number of tokens translated from 10 source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. Although the speeches are delivered orally (many times read out from written texts), they can be considered a translation rather than interpretation, since the proceedings are produced in the following way:[3]

1. The original speech is transcribed and minimally edited;
2. The text is sent to the speaker, who may edit it further;
3. The resulting text is translated into the other official languages.

The corpus is first tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary). The purpose of this is to make sure that the length of an article does not interfere with the classification. We thus obtain 2000 chunks of original English, and 200 chunks of translations from each of the ten source languages. We then generate POS tags for the tokenized texts. We use the UIUC CCG sentence segmentation tool[4] to detect sentence boundaries; and OpenNLP,[5] with the default MaxEnt tagger and Penn Treebank tagset, to tokenize the texts and induce POS tags.

We use the Weka toolkit (Hall et al., 2009) for classification; in all experiments, we use SVM (SMO) as the classification algorithm, with the default linear kernel. We employ ten-fold cross-validation[6] and report *accuracy* (percentage of chunks correctly classified). Since the classification task is binary and the training corpus is balanced, the baseline is 50%.

4 Hypotheses

We test several translation studies hypotheses. In this section we list each hypothesis, and describe how we model it in terms of the features used for classification. Feature design is a sophisticated process. In determining the feature set, the most important features must:

1. reflect frequent linguistic characteristics we would expect to be present in the two types of text;

[3] We are grateful to Emma Wagner, Vicki Brett, and Philip Cole (EU, Head of the Irish and English Translation Unit) for this information.

[4] `http://cogcomp.cs.illinois.edu/page/tools_view/2`, accessed 24 August 2012.

[5] `http://incubator.apache.org/opennlp/`, accessed 24 August 2012.

[6] In ten-fold cross validation, 90% of the annotated data are used for training, and the remaining 10% are used for testing. This process is repeated ten times, with different splits of the data, and the ten results are averaged. This guarantees the robustness of the evaluation, and minimizes the risk of over-fitting to the training data.

2. be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.; and

3. be easy to interpret, yielding insights regarding the differences between original and translated texts.

We focus on features that reflect structural properties of the texts, some of which have been used in previous works. We now define the features we explore in this work; for each feature, we provide a precise definition that facilitates replication of our results, as well as a hypothesis on its ability to distinguish between O and T, based on the translation studies literature.

When generating many of the features, we normalize the feature's value, v, by the number of tokens in the chunk, n: $v' = v \times 2000/n$. This balances the values over chunks that have slightly more or less than 2000 tokens each (recall that chunks respect sentence boundaries). Henceforth, when describing a normalized feature, we report v' rather than v. We also multiply the values of some features by some power of 10, rounding up the result to the nearest integer, in order to have a set of values that is easier to compare. This does not affect the classification results.

4.1 Simplification

Simplification refers to the process of rendering complex linguistic features in the source text into simpler features in the target text. Strictly speaking, this phenomenon can be studied only vis-à-vis the source text, since 'simpler' is defined here in reference to the source text, where, for example, the practice of splitting sentences or refraining from complex subordinations can be observed. And indeed, this is how simplification was first defined and studied in translation studies (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985). Baker (1993) suggests that simplification can be studied by comparing translated texts with non-translated ones, as long as both texts share the same domain, genre, time frame, etc. In a series of corpus-based studies, Laviosa (1998, 2002) confirms this hypothesis. Ilisei et al. (2010) and Ilisei and Inkpen (2011) train a classifier enriched by simplification features and bring further evidence for this universal in Romanian and Spanish.

We model the simplification hypothesis through the following features:[7]

Lexical Variety The assumption is that original texts are richer in terms of vocabulary than translated ones, as hypothesized by Baker (1993) and studied by Laviosa (1998). Lexical variety is known to be an unstable phenomenon which is highly dependent on corpus size (Tweedie and

[7]Of the features we define in this section, the first five were also implemented by Ilisei et al. (2010), who, in addition, added sentence depth (as the depth of the parse tree) and ambiguity (as the average number of senses per word).

Baayen, 1998). We therefore use three different *type-token ratio* (TTR) measures, following Grieve (2007), where V is the number of types and N is the number of tokens per chunk. All three versions consider punctuation marks as tokens.

1. V/N, magnified by order of 6.
2. $\log(V)/\log(N)$, magnified by order of 6.
3. $100 \times \log(N)/(1-V_1/V)$, where V_1 is the number of types occurring only once in the chunk.

Mean word length (in characters) We assume that translated texts use simpler words, in particular shorter ones. Punctuation marks are excluded from the tokens in this feature.

Syllable ratio We assume that simpler words are used in translated texts, resulting in fewer syllables per word. We approximate this feature by counting the number of vowel-sequences that are delimited by consonants or space in a word, normalized by the number of tokens in the chunk.

Lexical density This measure is also used by Laviosa (1998). The frequency of tokens that are *not* nouns, adjectives, adverbs or verbs. This is computed by dividing the number of tokens tagged with POS tags that do not open with J, N, R or V by the number of tokens in the chunk.

Mean sentence length Splitting sentences is a common strategy in translation, which is also considered a form of simplification. Baker (1993) renders it one of the universal features of 'simplification'. Long and complicated sentences may be simplified and split into short, simple sentences. Hence we assume that translations contain shorter sentences than original texts. We consider punctuation marks as tokens in the computation of this feature.

Mean word rank We assume that less frequent words are used more often in original texts than in translated ones. This is based on the observation of Blum-Kulka and Levenston (1983) that translated texts "make do with less words" and the application of this feature by Laviosa (1998). A theoretical explanation is provided by Halverson (2003): translators use more prototypical language, i.e., they "regress to the mean" (Shlesinger, 1989). To compute this, we use a list of 6000 English most frequent words,[8] and consider the *rank* of words (their position in the frequency-ordered list). The maximum rank is 5000 (since some words have equal ranks). We handle words that do not appear in the list in two different ways:

[8] `http://www.insightin.com/esl/`, accessed 24 August 2012.

1. Words not in this list are given a unique highest rank of 6000.
2. Words not in the list are ignored altogether.

Values (in both versions) are rounded to the nearest integer. All punctuation marks are ignored.

Most frequent words The normalized frequencies of the N most frequent words in the corpus. We define three features, with three different thresholds: $N = 5, 10, 50$. Punctuation marks are excluded.

4.2 Explicitation

Explicitation is the tendency to spell out in the target text utterances that are more implicit in the source. Like simplification, this 'universal' can be directly observed in T only in reference to O; if there is an implicit causal relation between two phrases in the source text and a cohesive marker such as *because* is introduced in target text, then it could be said with confidence that explicitation took place. But explicitation can also be studied by constructing a comparable corpus (Baker, 1993), and it is fair to assume that if there are many more cohesive markers in T than in O (in a well-balanced large corpus like EUROPARL), it could serve as an *indirect* evidence of explicitation.

Blum-Kulka (1986) develops and exemplifies this phenomenon in translations from Hebrew to English, and Øverås (1998) compiles a parallel bidirectional Norwegian-English and English-Norwegian corpus to provide further evidence for explicitation. Koppel and Ordan (2011) find that some of the prominent features in their list of function words are cohesive markers, such as *therefore*, *thus* and *consequently*.

The first three classifiers below are inspired by an example provided by Baker (1993, pp. 243-4), where the clause *The example of Truman was always present in my mind* is rendered into Arabic with a fairly long paragraph, which includes the following: *In my mind there was always the example of the American President Harry Truman, who succeeded Franklin Roosevelt....*

Explicit naming We hypothesize that one form of explicitation in translation is the use of a proper noun as a spelling out of a personal pronoun. We calculate the ratio of personal pronouns to proper nouns, both singular and plural, magnified by an order of 3. See also 'Pronouns', Section 4.5.

Single naming The frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor. This can be seen in an exaggerated form in the example above taken from Baker (1993, pp. 243-4). As a contemporary example, it is common to find

in German news (as of 2012) the single proper name *Westerwelle*, but in translating German news into another language, the translator is likely to add the first name of this person (*Guido*) and probably his role, too (*minister of foreign affairs*).

Mean multiple naming The average length (in tokens) of proper nouns (consecutive tokens tagged as Proper Nouns), magnified by an order of 3. The motivation for this feature is the same as above.

Cohesive markers Translations are known to excessively use certain *cohesive markers* (Blum-Kulka, 1986; Øverås, 1998; Koppel and Ordan, 2011). We use a list of 40 such markers, based on Koppel and Ordan (2011); see Appendix A.1. Each marker in the list is a feature, whose value is the frequency of its occurrences in the chunk.

4.3 Normalization

Translators take great efforts to standardize texts (Toury, 1995), or, in the words of (Baker, 1993, p. 244), they have "a strong preference for conventional 'grammaticality'". We include in this the tendency to avoid repetitions (Ben-Ari, 1998), the tendency to use a more formal style manifested in refraining from the use of contractions (Olohan, 2003), and the tendency to overuse fixed expressions even when the source text refrains, sometime deliberately, from doing so (Toury, 1980; Kenny, 2001).

We model normalization through the following features:

Repetitions We count the number of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk, and normalize by the number of tokens in the chunk. Inflections of the verbs *be* and *have* are excluded from the count since these verbs are commonly used as auxiliaries. This feature's values are magnified by an order of 3.

Contractions The ratio of contracted forms to their counterpart full form(s). If the full form has zero occurrences, its count is changed to 1. The list of contracted forms used for this feature is given in Appendix A.2.

Average PMI We expect original texts to use more collocations, and in any case to use them differently than translated texts. This hypothesis is based on Toury (1980) and Kenny (2001), who show that translations overuse highly associated words. We therefore use as a feature the average PMI (Church and Hanks, 1990) of all bigrams in the chunk. Given a bigram w_1w_2, its PMI is:

$$\log(^{freq(w_1w_2)}/_{freq(w_1)\times freq(w_2)})$$

Threshold PMI We compute the PMI of each bigram in a chunk, and count the (normalized) number of bigrams with PMI above 0.

4.4 Interference

Toury (1979) takes on the concept of *interlanguage* (Selinker, 1972) to define *interference* as a universal. Selinker (1972) coins the term in order to talk about the hybrid nature of the output of non-native speakers producing utterances in their second language. This output is heavily influenced by the language system of their first language. Translation is very similar in this sense, one language comes in close contact with another through transfer. In translation, however, translators habitually produce texts in their native tongue. Therefore, Toury (1979) advocates a descriptive study of interference not tainted, like in second language acquisition, by the view that the output reveals "ill performances" (production of grammatically incorrect structures). Interference operates on different levels, from transcribing source language words, through using loan translations, to exerting structural (morphological and syntactic for example) influence. This may bring about, as noted by Gellerstam (1986), a different distribution of elements in translated texts, which he calls 'translationese', keeping it as a pure descriptive term (cf. Santos (1995)).

We model interference as follows:

POS n-grams We hypothesize that different grammatical structures used in the different source languages interfere with the translations; and that translations have unique grammatical structure. Following Baroni and Bernardini (2006) and Kurokawa et al. (2009), we model this assumption by defining as features unigrams, bigrams and trigrams of POS tags. We add special tokens to indicate the beginning and end of each sentence, with the purpose of capturing specific POS-bigrams and POS-trigrams representing the beginnings and endings of sentences. The value of these features is the actual number of each POS n-gram in the chunk.

Character n-grams This feature is motivated by Popescu (2011). Other than yielding very good results, it is also language-type dependent. We hypothesize that grammatical structure manifests itself in this feature, and as in POS n-grams, the different grammatical structures used in the different source languages interfere with the translations. We also hypothesize that this feature captures morphological features of the language. These are actually three different features (each tested separately): unigrams, bigrams and trigrams of characters. They are computed similarly to the way POS n-grams are computed: by the frequencies of n-letter occurrences in a chunk, normalized by the chunk's size. Two special tokens are added to indicate the beginning and end

of each word, in order to properly handle specific word prefixes and suffixes. We do not capture cross-token character n-grams, and we exclude punctuation marks.

Prefixes and suffixes Character n-grams are an approximation of morphological structure. In the case of English, the little morphology expressed by the language is typically manifested as prefixes and suffixes. We therefore define a more refined variant of the character n-gram feature, focusing only on prefixes and suffixes. We use a list of such morphemes (see Appendix A.3) as features, simply counting the number of words in a chunk that begin or end with each of the prefixes/suffixes, respectively.

Contextual function words This feature is a variant of POS n-grams, where the n-grams can be anchored by specific (function) words. Koppel and Ordan (2011) use only function words for classification; we use the same list of words[9] in this feature (see Appendix A.4). This feature is defined as the (normalized) frequency of trigrams of function words in the chunk. In addition, we count trigrams consisting of two function words (from the same list) and one other word; in such cases, we replace the other word by its POS. In sum, we compute the frequencies in the chunk of triplets $\langle w_1, w_2, w_3 \rangle$, where at least two of the elements are functions words, and at most one is a POS tag.

Positional token frequency Writers have a relatively limited vocabulary from which to choose words to open or close a sentence. We hypothesize that the choices are subject to interference. Munday (1998) and Gries and Wulff (2012) study it on a smaller scale in translations from Spanish to English and in translations from English to German, respectively. The value of this feature is the normalized frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence. We exclude sentences shorter than five tokens. Punctuation marks are considered as tokens in this feature, and for this reason the three last positions of a sentence are considered, while only the first two of them are interesting for our purposes.

4.5 Miscellaneous

Finally, we define a number of features that cannot be naturally associated with any of the above hypotheses, but nevertheless throw light on the nature of translationese.

Function words We aim to replicate the results of Koppel and Ordan (2011) with this feature. We use the same list of function words (in

[9] We thank Moshe Koppel for providing us with the list of function words used in Koppel and Ordan (2011).

fact, some of them are content words, but they are all crucial for organizing the text; see the list in Appendix A.4) and implement the same feature. Each function word in the corpus is a feature, whose value is the normalized frequency of its occurrences in the chunk.

Pronouns Pronouns are function words, and Koppel and Ordan (2011) report that this subset is among the top discriminating features between O and T. We therefore check whether pronouns alone can yield a high classification accuracy. Each pronoun in the corpus is a feature, whose value is the normalized frequency of its occurrences in the chunk. The list of pronouns is given in Appendix A.5.

Punctuation Punctuation marks organize the information within sentence boundaries and to a great extent reduce ambiguity; according to the explicitation hypothesis, translated texts are less ambiguous (Blum-Kulka, 1986) and we assume that this tendency will manifest itself in the (different) way in which translated texts are punctuated. We focus on the following punctuation marks: *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .* Apostrophes used in contracted forms are retained. Following Grieve (2007), we define three variants of this feature:

1. The normalized frequency of each punctuation mark in the chunk.
2. A non-normalized notion of frequency: $n/tokens$, where n is the number of occurrences of a punctuation mark; and *tokens* is the actual (rather than normalized) number of tokens in the chunk. This value is magnified by an order of 4.
3. n/p, where p is the total number of punctuations in the chunk; and n as above. This value is magnified by an order of 4.

Ratio of passive forms to all verbs We assume that English original texts tend to use the passive form more excessively than translated texts, due to the fact that the passive voice is more frequent in English than in some other languages (cf. Teich (2003) for German-English). If an active voice is used in the source language, translators may prefer not to convert it to the passive. Passives are defined as the verb *be* followed by the POS tag *VBN* (past participle). We calculate the ratio of passive verbs to all verbs, and magnified it by an order of 6.

As a "sanity check", we use two other features: token unigrams and token bigrams. Each unigram and bigram in the corpus constitutes a specific feature, as in Baroni and Bernardini (2006). The feature's value is its frequency in the chunk (again, normalized). For bigrams we add special markers of the edges of the sentences as described for POS-n-grams. We assume that different languages use different content words in varying frequencies in translated and non-translated texts. We expect these two

features to yield conclusive results (well above 90% accuracy), while token bigrams are expected to yield somewhat better results than token unigrams. These features are highly content-dependent, and are therefore of no empirical significance; they are only used as an upper bound for our other features, and to emphasize the validity of our methodology: we expect very high accuracy of classification with these features.

5 Results

We implemented all the features discussed in the previous section as classifiers and used them for classifying held-out texts in a ten-fold cross-validation scenario, as described in Section 3. The results of the classifiers are reported in Table 2 in terms of the accuracy of classifying the test set.

As a sanity check, we also report the accuracy of the content-dependent classifiers. As mentioned above, these are expected to produce highly-accurate classifiers, but teach us very little about the features of translationese. As is evident from Table 3, this is indeed the case.

For the sake of completeness, we note that it is possible to achieve very high classification accuracy even with a much narrower feature space. Some of the more complex feature sets have hundreds, or even thousands of features. In such cases, most features contribute very little to the task. To emphasize this, we take only the top 300 most frequent features. For example, rather than use *all* possible POS trigrams for classification, we only use the 300 most frequent sequences as features. Table 4 lists the classification results in this case. Evidently, the results are almost as high as when using *all* features.

Our main objective, however, is not to produce the best-performing classifiers. Rather, it is to understand what the classifiers can reveal about the nature of the differences between O and T. The following section thus analyses the results.

6 Analysis

6.1 Simplification

Laviosa (1998, 2002) studied the simplification hypothesis extensively. Some features pertaining to simplification are also mentioned by Baker (1993). The four main features and partial findings pertain to mean sentence length, type-token ratio, lexical density and overrepresentation of highly frequent items. Lexical density fails altogether to predict the status of a text, being nearly on chance level (53% accuracy). Interestingly, while mean sentence length is much above chance level (65%), the results are contrary to common assumptions in Translation Studies. According to the simplification

Category	Feature	Accuracy (%)
Simplification	TTR (1)	72
	TTR (2)	72
	TTR (3)	76
	Mean word length	66
	Syllable ratio	61
	Lexical density	53
	Mean sentence length	65
	Mean word rank (1)	69
	Mean word rank (2)	77
	N most frequent words	64
Explicitation	Explicit naming	58
	Single naming	56
	Mean multiple naming	54
	Cohesive Markers	81
Normalization	Repetitions	55
	Contractions	50
	Average PMI	52
	Threshold PMI	66
Interference	POS unigrams	90
	POS bigrams	97
	POS trigrams	98
	Character unigrams	85
	Character bigrams	98
	Character trigrams	100
	Prefixes and suffixes	80
	Contextual function words	100
	Positional token frequency	97
Miscellaneous	Function words	96
	Pronouns	77
	Punctuation (1)	81
	Punctuation (2)	85
	Punctuation (3)	80
	Ratio of passive forms to all verbs	65

Table 2: Classification results

hypothesis, T sentences are simpler (i.e., shorter), but as Figure 1 shows, the contrary is the case. We computed the mean sentence length in eleven 400,000-word texts, one of them original English, and the other translated from ten source languages (this is the same corpus on which we run the classification). Only three translations (from Swedish, Finnish and Dutch) have a lower mean sentence length than original English, and on average O sentences are 2.5 tokens shorter. Whereas this result may pertain only to certain language pairs or certain genres, this alleged "translation universal" is definitely not universal. Moreover, it may actually be an instance of the

Category	Feature	Accuracy (%)
Sanity	Token unigrams	100
	Token bigrams	100

Table 3: Classification results, "sanity check" classifiers

Category	Feature	Accuracy
Interference	POS bigrams	96
	POS trigrams	96
	Character bigrams	95
	Character trigrams	96
	Positional token frequency	93

Table 4: Classification results, top-300 features only

interference hypothesis, where sentence length in the target language reflects its length in the source language. This, however, should be studied under a parallel corpus setting, and is beyond the scope of this work.

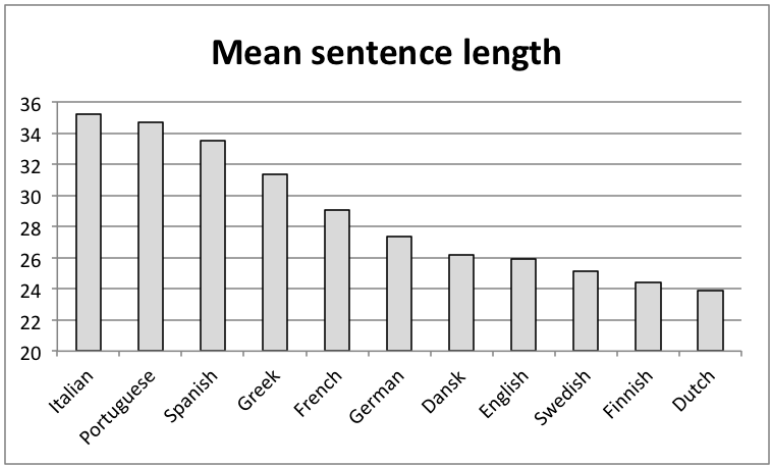

Figure 1: Mean sentence length according to 'language'

The first two TTR measures perform relatively well (72% accuracy), and the indirect measures of lexical variety (mean word length and syllable ratio) are above chance level (66% and 61% accuracy, respectively). Following Holmes (1992) we experiment with more sophisticated measures of lexical variety. The best performing one is the one that takes into account *hapax legomena*, words that occur only once in a text. This variant of TTR (3) yields 76% accuracy. One important trait of *hapax legomena* is that as op-

posed to type-token ratio they are not so dependent on corpus size (Baayen, 2001).

Another interesting classifier with relatively good results, in fact, the best performing of all simplification features (77% accuracy), is mean word rank. This feature is closely related to the feature studied by Laviosa (1998) (n top words) with two differences: (1) our list of frequent items is much larger, and (2) we generate the frequency list not from the corpora under study but rather from an external much larger reference corpus. In contrast, the design that follows Laviosa (1998) more strictly (N most frequent words) has a lower predictive power (64%).

6.2 Explicitation

The three classifiers we design to check this hypothesis (explicit naming, single naming and mean multiple naming) do not exceed 58% classification accuracy. On the other hand, following Blum-Kulka (1986) and Koppel and Ordan (2011), we build a classifier that uses 40 cohesive markers and achieve 81% accuracy in telling O from T; such cohesive markers are far more frequent in T than in O. For example, *moreover*, *thus* and *besides* are used 17.5, 4, and 3.8 times more frequently (respectively) in T than in O.

6.3 Normalization

None of these features perform very well. Repetitions and contractions are rare in EUROPARL and in this sense the corpus may not be suited for studying these phenomena. The repetition-based classifier yields 55% accuracy and the contraction-based classifier performs at chance level (50%).

One of the classifiers that checks PMI, designed to pick on highly associated words and therefore attesting to many fixed expressions, performs considerably better, namely 66% accuracy. This measure counts the number of associated bigrams whose PMI is above 0. As Figure 2 shows, English has far more highly associated bigrams than translations. If we take the word form *stand*, for example, then at the top of the list we normally get highly associated words, some of which are fixed expressions, such as *stand idly*, *stand firm*, *stand trial*, etc. There are considerably more highly associated pairs like these in O; conversely, this also means that there are more poorly associated pairs in T, such as the bigram *stand unamended*. This finding contradicts the case studies elaborated on in Toury (1980); Kenny (2001). It should be noted, however, that they discuss cases operating under particular scenarios, whereas we check this phenomenon more globally, completely unbiased towards any scenario whatsoever. The finding is robust but it is oblivious to the particulars of subtle cases.

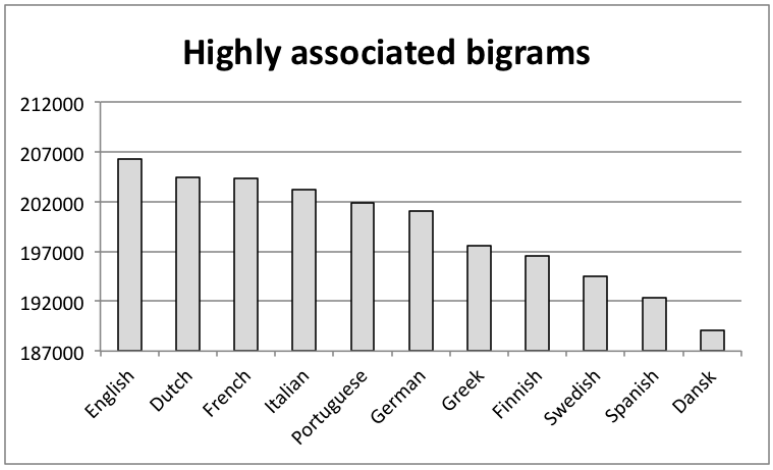

Figure 2: Number of bigrams whose PMI is above threshold according to 'language'

6.4 Interference

The interference-based classifiers are the best performing ones. Most of them perform above 90%. In this sense we can say that interference is the most robust phenomenon typifying translations. However, we note that some of the features are somewhat coarse and may reflect some corpus-dependent characteristics. For example, in the character n-grams we notice that some of the top features in O include sequences that are 'illegal' in English and obviously stem from foreign names, such as the following letter bigrams: *Haa*rder and *Maa*t or *Gazp*ron. To offset this problem we use only the top 300 features in several of the classifiers, with a minor effect on the results.

The n-gram findings are consistent with Popescu (2011) in that we also find they catch on both affixes and function words: for example, typical trigrams in O are *-ion* and *all* whereas typical to T are *-ble* and *the*. As opposed to Popescu (2011) we reduced the feature space without the need to look at the original texts; Popescu (2011) looked for sequences of n-grams in the target language that also appear in the source texts, thereby eliminating mostly proper nouns. However, this method can be applied only to language pairs that use similar alphabet and orthography conventions. Using only the 300 most frequent features results in a drop in accuracy of up to 4%. Furthermore, restricting the space of features to only 38 prefixes and 34 suffixes, a much narrower domain than the set of all character bi-grams, for example, still yields 80% accuracy. Evidently, original and translated texts differ greatly in the way they use these affixes.

Different English affixes were imported from different languages, and this is reflected in our findings. The prefix *mono-*, a marker of translated

language, is much more frequent in Greek than any other language. The suffix *-ible*, originating in Latin, is much more common in all the Romance languages, which are "clustered together" around this feature, compared to English. Last, the suffix *-ize*, originating in Latin, is highly frequent in original English, less frequent in the Romance languages, and even less in the other languages. Further study, backed by a sound historical linguistics perspective, may determine how such parameters affect transnational choices between language, taking into account their distance from each other.

Part-of-speech trigrams is an extremely cheap and efficient classifier. The feature space is not too big, and the results are robust. A good discriminatory feature typifying T is, for example, the part-of-speech trigram of modal + verb base form + verb past particle, as in the highly frequent phrases in the corpus *must be taken*, *should be given* and *can be used*; as can be seen in Figure 3, it typifies more prominently translations from phylogenetically distant languages, such as Finnish, but original English is down the list, regardless of T's source language.

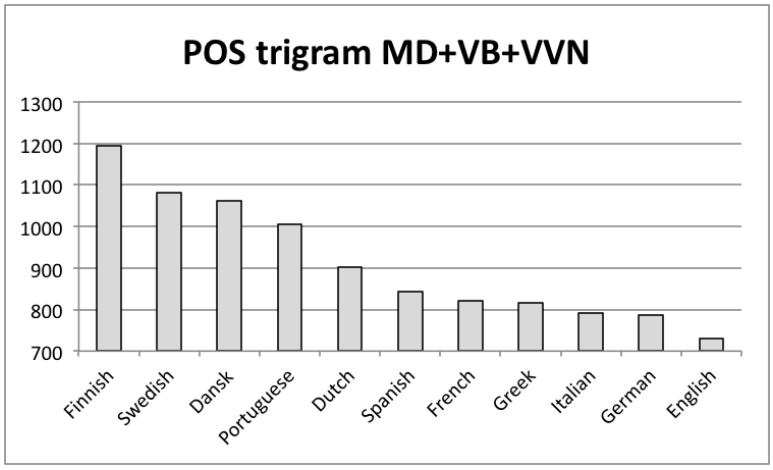

Figure 3: The average number of the POS trigram modal + verb base form + participle in O and ten Ts

Moving now to positional token frequency, we report on three variations of this classifier with different degrees of accuracy (reported in brackets): taking into account all the tokens that appear in these positions (97%), using only the 300 most frequent tokens (93%) and finally only the 50 most frequent tokens (82%). The last is the most abstract, picking almost exclusively on function words. The second most prominent feature typifying O is sentences opening with the word 'But'. In fact, there are 2.25 times more cases of such sentences in O. In English there is a long prescriptive tradition forbidding writers to open a sentence with 'But', and although this 'decree' is questioned and even mocked at (Garner, 2003), the question

whether it is considered a good style is a common question posted on Internet forums dealing with English language use. Translators have been known to be conservative in their lexical choices (Kenny, 2001), and the underuse of 'But'-opening sentences is yet another evidence for this tendency. As opposed to other features in positional token frequency, this is *not* a case of interference but rather a tendency to (over-)abide to norms of translation, i.e., *standardization* (Toury, 1995).

6.5 Miscellaneous

In this category we include several classifiers whose features do not fall under a clear-cut theoretical category discussed by translation theorists. The function words classifier replicates Koppel and Ordan (2011) and despite the good performance (96% accuracy) it is not very meaningful theoretically. One of its subsets, a list of 25 pronouns, reveals an interesting phenomenon: subject pronouns, like *I*, *he* and *she* are prominent indicators of O, whereas virtually all reflexive pronouns (such as *itself*, *himself*, *yourself*) typify T. The first phenomenon is probably due to the fact that pronouns are much more frequent in T (about 1.25 more frequent) and a fine-tuned analysis of the distribution of pronouns in each sub-corpus normalized by the number of pronouns is beyond the scope of this study; the high representation of reflexive pronouns is probably due to interference from the source languages. The accuracy of classifying by pronouns alone is 77%.

The accuracy of a classifier based on the ratio of passive verbs is much above chance level, yet not a very good predictor by itself (65%). T has about 1.15 times more passive verbs, and it is highly dependent on the source language from which T stems: original English is down the list, right after the Romance languages and Greek, and from the top down: Danish, Swedish, Finnish, Dutch and German.

We experiment with three different classifiers based on punctuation marks as a feature set. The mark '.' (actually indicating sentence length) is a strong feature of O and the mark ',' is a strong marker of T. In fact, using only these two features we achieve 79% accuracy. Parentheses are very typical to T, indicating explicitation. A typical example is the following: *The Vlaams Blok ('Flemish Block') opposes the patentability of computer-implemented inventions...* Last, we find that exclamation marks are on average much more common in original English (1.25 times more frequent). Translations from three source languages, however, have more exclamation marks than original English: German, Italian and French. Translations from German use many more exclamation marks, 2.76 (!!!) times more than original English.

7 Conclusion

Machines can easily identify translated texts. Identification has been successfully performed for very different data sets and genres, including parliamentary proceedings, literature, news and magazine writing, and it works well across many source and target languages (with the exception of literary Polish, see Rybicki (2012)). But text classification is a double-edged sword. Consider how easily the classifier teases apart O from T based on letter bigrams: 98% accuracy, with a slight drop to 95% when only the top 300 most frequent letter bigrams are used. It is considerably better than the performance achieved by professional humans (Tirkkonen-Condit, 2002; Baroni and Bernardini, 2006). We then find that the letter sequence *di* is among the best discriminating features between O and T, as it is about 16% more frequent in T than in O; but it does not teach us much about T, and we cannot interpret this finding. Furthermore, text classification is highly dependent on the genres and domains, and cross-corpus classification ('scalability') is notoriously hard (Argamon, 2011).

We addressed the first problem by designing linguistically informed features. For example, enhancing letter n-grams to trigrams already revealed some insights about morphological traits of T. The second problem calls for future research. Recall that we were unable to replicate the results reported by Olohan (2003), simply because contractions are a rarity in EUROPARL and therefore 'normalizing' them is even a rarer event. That *translationese* is dependent on genre is suggested and studied in various works (Steiner, 1998; Reiss, 1989; Teich, 2003).

This point is much related to one of our main conclusions: the universal claims for translation should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages. The best performing features in our study are those that attest to the 'fingerprints' of the source on the target, what has been called "source language shining through" (Teich, 2003). This is not to say that there are no features which operate "irrespective of source language" (like cohesive markers in EUROPARL), but the best evidence for translationese, the one that has the best predictive power, is related to interference, and interference by its nature is a pair-specific phenomenon. Note that mean sentence length, which we included in 'simplification', has been purported to be a trait of translationese regardless of source language, but turned out to be very much dependent on the source language, and in particular, contrary to previous assumptions, sentence length turned out to be shorter in O. This can indeed be shown in a well-balanced comparable corpus, ideally from as many source languages as possible and, when possible, typologically distant ones.

Another caveat is related to comparable corpora in general. Olohan and Baker (2000) report that there are less omissions of optional reporting *that*

in T, as in *I know (that) he'd never get here in time*. This is, according to the authors, a case of *explicitation*, i.e., replacing a zero-connective with a that-connective to avoid ambiguity. Pym (2008) raises the following question: how do we know that this finding is not due to interference? What if the T component of this corpus consists of source languages in which the *that*-connective is obligatory and therefore it is just "shining through" to the target text? We cast the same doubt on some of our findings. The under-representation of sentences opening with *But* in T are probably due to normalization, but without reference to the source texts we will never be sure. With this kind of corpus — a comparable corpus — we can settle the ontological question (T is different from O across many dimensions suggested by translation scholars), but we are left with an epistemological unease: given our tools and methodology we do not know for sure what part of the findings is a mere result of source influence on the target text (interference), and what part is inherent to the work of translators (simplification, normalization, and explicitation). We leave this question for future studies.

References

Omar S. Al-Shabab. *Interpretation and the language of translation: creativity and conventions in translation*. Janus, Edinburgh, 1996.

Shlomo Argamon. Book review of *Scalability Issues in Authorship Attribution*, by Kim Luyckx. *Literary and Linguistic Computing*, 27(1):95–97, 12 2011.

R. Harald Baayen. *Word Frequency Distributions*. Text, speech, and language technology. Kluwer Academic, 2001. ISBN 9780792370178.

Mona Baker. 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, 1993.

Marco Baroni and Silvia Bernardini. 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, September 2006. URL `http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1`.

Nitza Ben-Ari. The ambivalent case of repetitions in literary translation. Avoiding repetitions: A "universal" of translation? *Meta*, 43(1):68–78, 1998.

Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and inter-*

cultural communication Discourse and cognition in translation and second language acquisition studies, volume 35, pages 17–35. Gunter Narr Verlag, 1986.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. *Language Learning*, 28(2):399–416, December 1978.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, *Strategies in Interlanguage Communication*, pages 119–139. Longman, 1983.

Andrew Chesterman. Beyond the particular. In A. Mauranen and P. Kujamäki, editors, *Translation universals: Do they exist?*, pages 33–50. John Benjamins, 2004.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. ISSN 0891-2017.

Sari Eskola. Untypical frequencies in translated language. In A. Mauranen and P. Kujamäki, editors, *Translation universals: Do they exist?*, pages 83–99. John Benjamins, 2004.

William Frawley. Prolegomenon to a theory of translation. In William Frawley, editor, *Translation. Literary, Linguistic and Philosophical Perspectives*, pages 159–175. University of Delaware Press, Newark, 1984.

Bryan A. Garner. On beginning sentences with But. *Michigan Bar Journal*, 48:43–4, 2003.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

Stefan Th. Gries and Stefanie Wulff. Regression analysis in translation studies. In Michael P. Oakes and Meng Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies*, Studies in Corpus Linguistics 51, pages 35–52. John Benjamins, Philadelphia, 2012.

Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://dx.doi.org/10.1145/1656274.1656278`.

Sandra Halverson. The cognitive basis of translation universals. *Target*, 15 (2):197–241, 2003.

David I. Holmes. A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society*, 155(1):91–120, 1992.

Iustina Ilisei. *A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models*. PhD thesis, University of Wolverhampton, Wolverhampton, UK, February 2013. URL `http://clg.wlv.ac.uk/papers/ilisei-thesis.pdf`.

Iustina Ilisei and Diana Inkpen. Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1-2), 2011.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL `http://dx.doi.org/10.1007/978-3-642-12116-6`.

Dorothy Kenny. *Lexis and creativity in translation: a corpus-based study*. St. Jerome, 2001. ISBN 9781900650397.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86. AAMT, 2005. URL `http://mt-archive.info/MTS-2005-Koehn.pdf`.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1132`.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, Jan 2009. ISSN 1532-2882. doi: 10.1002/asi.v60:1. URL `http://dx.doi.org/10.1002/asi.v60:1`.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88, 2009.

Sara Laviosa. Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta*, 43(4):557–570, December 1998.

Sara Laviosa. *Corpus-based translation studies: theory, findings, applications*. Approaches to translation studies. Rodopi, 2002. ISBN 9789042014879.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1034`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France, April 2012a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E12-1026`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825, December 2012b. URL `http://dx.doi.org/10.1162/COLI_a_00111`.

A. Mauranen and P. Kujamäki, editors. *Translation universals: Do they exist?* John Benjamins, 2004.

Anna Mauranen. Universals tendencies in translation. In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: the linguist and the translator*, pages 32–48. Multilingual Matters, Clevedon, Buffalo and Toronto, 2008.

Jeremy Munday. A computer-assisted approach to the analysis of translation shifts. *Meta*, 43(4):542–556, 1998.

Maeve Olohan. How frequent are the contractions? A study of contracted forms in the translational English corpus. *Target*, 15(1):59–89, 2003.

Maeve Olohan and Mona Baker. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1(2):141–158, 2000.

Lin Øverås. In search of the third code: An investigation of norms in literary translation. *Meta*, 43(4):557–570, 1998.

Marius Popescu. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *Proceedings of RANLP-2011*, pages 634–639, 2011.

Anthony Pym. On Toury's laws of how translators translate. In Anthony Pym, Miriam Shlesinger, and Daniel Simeoni, editors, *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*, Benjamins translation library: EST subseries, pages 311–328. John Benjamins, 2008. ISBN 9789027216847.

Katherine Reiss. Text types, translation types and translation assessment. In Andrew Chesterman, editor, *Readings in translation theory*, pages 105–115. Oy Finn Lectura Ab, Helsinki, 1989.

Jan Rybicki. The great mystery of the (almost) invisible translator: Stylometry in translation. In Michael P. Oakes and Meng Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies*, Studies in Corpus Linguistics 51, pages 231–248. John Benjamins, Philadelphia, 2012.

Diana Santos. On grammatical translationese. In Kimmo Koskenniemi, editor, *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*, pages 29–30, 1995.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002. ISSN 0360-0300. doi: 10.1145/505282.505283. URL `http://doi.acm.org/10.1145/505282.505283`.

Larry Selinker. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1–4):209–232, 1972.

Miriam Shlesinger. Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum. Master's thesis, Tel Aviv University, Faculty of the Humanities, Department of Poetics and Comparative Literature, 1989.

Erich Steiner. A register-based translation evaluation: An advertisement as a case in point. *Target*, 10(2):291–318, 1998.

Elke Teich. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, 2003.

Sonja Tirkkonen-Condit. Translationese: A myth or an empirical fact? *Target*, 14(2):207–220, 2002.

Gideon Toury. Interlanguage and its manifestations in translation. *Meta*, 24(2):223–231, 1979.

Gideon Toury. *In Search of a Theory of Translation.* The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.

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

Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

Hans van Halteren. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 937–944, 2008. ISBN 978-1-905593-44-6. URL `http://www.aclweb.org/anthology/C08-1118`.

Ria Vanderauwerea. *Dutch novels translated into English: the transformation of a 'minority' literature.* Rodopi, Amsterdam, 1985.

A Lists of words

A.1 Cohesive markers

We use the following list of words as cohesive markers: *as for, as to, because, besides, but, consequently, despite, even if, even though, except, further, furthermore, hence, however, in addition, in conclusion, in other words, in spite, instead, is to say, maybe, moreover, nevertheless, on account of, on the contrary, on the other hand, otherwise, referring to, since, so, the former, the latter, therefore, this implies, though, thus, with reference to, with regard to, yet, concerning.*

A.2 Contracted forms

We use the following list of contracted forms and their expansions: *i'm: i am, it's: it is, it has, there's: there is, there has, he's: he is, he has, she's: she is, she has, what's: what is, what has, let's: let us, who's: who is, who has, where's: where is, where has, how's: how is, how has, here's: here is, i'll: i will, you'll: you will, she'll: she will, he'll: he will, we'll: we will, they'll: they will, i'd: i would, i had, you'd: you would, you had, she'd: she would, she had, he'd: he would, he had, we'd: we would, we had, they'd: they would, they had, i've: i have, you've: you have, we've: we have, they've: they have, who've: who have, would've: would have, should've: should have, must've: must have, you're: you are, they're: they are, we're: we are, who're: who are, couldn't: could not, can't: cannot, wouldn't: would not, don't: do not, doesn't: does not, didn't: did not.*

A.3 Prefixes and suffixes

We use the following list of prefixes: *a*, *an*, *ante*, *anti*, *auto*, *circum*, *co*, *com*, *con*, *contra*, *de*, *dis*, *en*, *ex*, *extra*, *hetero*, *homo*, *hyper*, *il*, *im*, *in*, *inter*, *intra*, *ir*, *macro*, *micro*, *mono*, *non*, *omni*, *post*, *pre*, *pro*, *sub*, *syn*, *trans*, *tri*, *un*, *uni* and the following list of suffixes: *able*, *acy*, *al*, *ance*, *ate*, *dom*, *en*, *ence*, *er*, *esque*, *ful*, *fy*, *ible*, *ic*, *ical*, *ify*, *ious*, *ise*, *ish*, *ism*, *ist*, *ity*, *ive*, *ize*, *less*, *ment*, *ness*, *or*, *ous*, *ship*, *sion*, *tion*, *ty*, *y*.

A.4 Function words

We use the following list of function words: *a*, *about*, *above*, *according*, *accordingly*, *actual*, *actually*, *after*, *afterward*, *afterwards*, *again*, *against*, *ago*, *ah*, *ain't*, *all*, *almost*, *along*, *already*, *also*, *although*, *always*, *am*, *among*, *an*, *and*, *another*, *any*, *anybody*, *anyone*, *anything*, *anywhere*, *are*, *aren't*, *around*, *art*, *as*, *aside*, *at*, *away*, *ay*, *back*, *be*, *bear*, *because*, *been*, *before*, *being*, *below*, *beneath*, *beside*, *besides*, *better*, *between*, *beyond*, *bid*, *billion*, *billionth*, *both*, *bring*, *but*, *by*, *came*, *can*, *can't*, *cannot*, *canst*, *certain*, *certainly*, *come*, *comes*, *consequently*, *could*, *couldn't*, *couldst*, *dear*, *definite*, *definitely*, *despite*, *did*, *didn't*, *do*, *does*, *doesn't*, *doing*, *don't*, *done*, *dost*, *doth*, *doubtful*, *doubtfully*, *down*, *due*, *during*, *e.g.*, *each*, *earlier*, *early*, *eight*, *eighteen*, *eighteenth*, *eighth*, *eighthly*, *eightieth*, *eighty*, *either*, *eleven*, *eleventh*, *else*, *enough*, *enter*, *ere*, *erst*, *even*, *eventually*, *ever*, *every*, *everybody*, *everyone*, *everything*, *everywhere*, *example*, *except*, *exeunt*, *exit*, *fact*, *fair*, *far*, *farewell*, *few*, *fewer*, *fifteen*, *fifteenth*, *fifth*, *fifthly*, *fiftieth*, *fifty*, *finally*, *first*, *firstly*, *five*, *for*, *forever*, *forgo*, *forth*, *fortieth*, *forty*, *four*, *fourteen*, *fourteenth*, *fourth*, *fourthly*, *from*, *furthermore*, *generally*, *get*, *gets*, *getting*, *give*, *go*, *good*, *got*, *had*, *has*, *hasn't*, *hast*, *hath*, *have*, *haven't*, *having*, *he*, *he'd*, *he'll*, *he's*, *hence*, *her*, *here*, *hers*, *herself*, *him*, *himself*, *his*, *hither*, *ho*, *how*, *how's*, *however*, *hundred*, *hundredth*, *i*, *i'd*, *i'm*, *i've*, *if*, *in*, *indeed*, *instance*, *instead*, *into*, *is*, *isn't*, *it*, *it'd*, *it'll*, *it's*, *its*, *itself*, *last*, *lastly*, *later*, *less*, *let*, *let's*, *like*, *likely*, *many*, *matter*, *may*, *maybe*, *me*, *might*, *million*, *millionth*, *mine*, *more*, *moreover*, *most*, *much*, *must*, *mustn't*, *my*, *myself*, *nay*, *near*, *nearby*, *nearly*, *neither*, *never*, *nevertheless*, *next*, *nine*, *nineteen*, *nineteenth*, *ninetieth*, *ninety*, *ninth*, *ninthly*, *no*, *nobody*, *none*, *noone*, *nor*, *not*, *nothing*, *now*, *nowhere*, *o*, *occasionally*, *of*, *off*, *oft*, *often*, *oh*, *on*, *once*, *one*, *only*, *or*, *order*, *other*, *others*, *ought*, *our*, *ours*, *ourselves*, *out*, *over*, *perhaps*, *possible*, *possibly*, *presumable*, *presumably*, *previous*, *previously*, *prior*, *probably*, *quite*, *rare*, *rarely*, *rather*, *result*, *resulting*, *round*, *said*, *same*, *say*, *second*, *secondly*, *seldom*, *seven*, *seventeen*, *seventeenth*, *seventh*, *seventhly*, *seventieth*, *seventy*, *shall*, *shalt*, *shan't*, *she*, *she'd*, *she'll*, *she's*, *should*, *shouldn't*, *shouldst*, *similarly*, *since*, *six*, *sixteen*, *sixteenth*, *sixth*, *sixthly*, *sixtieth*, *sixty*, *so*, *soever*, *some*, *somebody*, *someone*, *something*, *sometimes*, *somewhere*, *soon*, *still*, *subsequently*, *such*,

sure, tell, ten, tenth, tenthly, than, that, that's, the, thee, their, theirs, them, themselves, then, thence, there, there's, therefore, these, they, they'd, they'll, they're, they've, thine, third, thirdly, thirteen, thirteenth, thirtieth, thirty, this, thither, those, thou, though, thousand, thousandth, three, thrice, through, thus, thy, till, tis, to, today, tomorrow, too, towards, twas, twelfth, twelve, twentieth, twenty, twice, twill, two, under, undergo, underneath, undoubtedly, unless, unlikely, until, unto, unusual, unusually, up, upon, us, very, was, wasn't, wast, way, we, we'd, we'll, we're, we've, welcome, well, were, weren't, what, what's, whatever, when, whence, where, where's, whereas, wherefore, whether, which, while, whiles, whither, who, who's, whoever, whom, whose, why, wil, will, wilst, wilt, with, within, without, won't, would, wouldn't, wouldst, ye, yes, yesterday, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves.

A.5 Pronouns

We use the following list of pronouns: *he, her, hers, herself, him, himself, i, it, itself, me, mine, myself, one, oneself, ours, ourselves, she, theirs, them, themselves, they, us, we, you, yourself.*