Improving Statistical Machine Translation by Automatic Identification of Translationese

Naama Twitto-Shmuel

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa
Faculty of Social Sciences
Department of Computer Science

November, 2013

Improving Statistical Machine Translation by Automatic Identification of Translationese

By: Naama Twitto-Shmuel

Supervised By: Prof. Shuly Wintner and
Dr. Noam Ordan

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa
Faculty of Social Sciences
Department of Computer Science

November, 2013

Approved by: _____ Date:_____
(supervisor)

Approved by: _____ Date:_____
(supervisor)

Approved by: _____ Date:_____
(Chairman of M.A. Committee)

Acknowledgements

Great many thanks to my advisors, Prof. Shuly Wintner and Dr. Noam Ordan, for all I have learned from them and for their continuous help and support in all the stages of this thesis. Their help in processing, analysing, and assessing insights from the raw results was invaluable to the progress of this work. I would also like to thank them for helping me to shape the thesis to its final form.

During our work together I have studied through them the important and useful domain of Statistical Machine Translation, from the basics to advanced topics and best practices of this domain. I have learned how to carry out scientific research, experiments, and how to consolidate the results. I also practised the expertise of scientific writing, and how to communicate my results. For all of that I would like to thank my advisors.

Special thanks to Dr. Gennadi Lembersky and Vered Volansky for providing me important information, each on his work. I would like to thank them for helping me to reproduce their results, in the first steps of my work. This reproduction ability was crucial as my research is partially based on their results.

I would like to thank Haifa university and the department of Computer Science for providing me the background conditions to carry out this research, and for being a warm environment as well. Finally, I thank the Israeli Ministry of Science and Technology for financially supporting this research.

Contents

Improving Statistical Machine Translation by Automatic Identification of Translationese

Naama Twitto-Shmuel

Abstract

Translated texts are so markedly different from original ones that text classification techniques can be used to tease them apart. Furthermore, it was shown that awareness to these differences can significantly improve statistical machine translation. A prerequisite for these improvements, however, is meta-information on the ontological status of texts (original or translated); such meta-information is typically unavailable. In this work we set out to overcome this limitation by incorporating the predictions of translation classifiers in machine translation. Specifically, we show that the predictions of machine-learning-based classifiers can be as good as meta-information on the status of the texts. First, when a monolingual corpus in the target language is given, to be used for constructing a language model, we show that predicting the translated portions of the corpus, and using only them for the language model, is as good as using the entire corpus. Second, we show that automatically identifying the portions of a parallel corpus that are translated in the direction of the translation task, and using only them for the translation model, is as good as using the entire corpus. We thus propose a way to construct much smaller language- and translation-models that are as good as ones based on much larger corpora. We present results from several language pairs, indicating that the approach we advocate is robust and general.

List of Tables

1 Introduction

Research in Translation Studies suggests that translated texts are markedly different from original texts, constituting a genre, or a dialect, known as *Translationese* (Gellerstam, 1986). Awareness to translationese can significantly improve statistical machine translation (SMT). Kurokawa et al. (2009) showed that French-to-English SMT systems whose translation models were constructed from human translations from French to English yielded better translation quality than ones created from translations in the other direction; the same holds for SMT systems translating English to French. These results were corroborated by Lembersky et al. (2012a, 2013), who showed that translation models can be adapted to translationese, thereby improving the quality of SMT even further. Moreover, awareness to translationese also benefits the *language* models used for SMT: Lembersky et al. (2011, 2012b) showed that language models complied from translated texts better fit the reference sets in term of perplexity, and SMT systems that are constructed from these language models perform much better than those constructed from original texts.

To benefit from these results, however, one has to know whether the (monolingual and parallel) texts used for training SMT systems are original or translated. Such meta-information is typically unavailable. In this work we set out to overcome this limitation: instead of using meta-information about the ontological status of texts (original vs. translated), we use automatic classifiers to predict this status. When a monolingual corpus in the target language is given for constructing a language model for SMT, we show that automatically identifying the portions of the corpus that are translated, and using only those predicted portions for the language model, is as good as using the entire corpus. Similarly, when a parallel corpus is given, we show that automatically identifying the portions of the corpus that are translated in the direction of the translation task, and using only them for training the translation model, is again as good as using the entire corpus. We present results from several language pairs, indicating that the approach we advocate is not language-pair-specific. We also conduct cross-corpus evaluations that demonstrate the robustness of the approach.

The main contribution of this work, then, is a general approach that, provided labeled data for training classifiers, can be applied to *any* corpus before it is used for constructing SMT systems: to build a language model from a monolingual corpus, use classifiers to predict the ontological status of texts in the corpus, and select only those texts that are

predicted as translated. To construct the translation model from a parallel corpus, predict the translation direction of the text, and select only bitexts translated in the same direction as the direction of the SMT task. This results in SMT systems that are as good as (or even better than) those which use the entire corpora, but rely on significantly smaller language and translation models.

A secondary contribution is a practical investigation of the utility for this task of various classifiers (which differ in the features they use to represent texts). Several works employed machine-learning-based text-classification methods for identifying translationese: e.g., Baroni and Bernardini (2006) in Italian, Ilisei et al. (2010) in Spanish, Koppel and Ordan (2011) and Volansky et al. (Forthcoming) in English, and Avner (2013) in Hebrew. In particular, Volansky et al. (Forthcoming) investigated as many as 32 different feature sets; we show in this work that only some of these features are indeed effective for the task at hand.

We briefly review related work in Section 2. Section 3 describes our methodology and experimental setup. In Section 4 we detail the experiments and their results. An analysis of the results is provided in Section 5. We conclude with suggestions for future research.

2 Related work

SMT is the prevalent paradigm in contemporary machine translation. It is based on the *noisy channel* model of Brown et al. (1990): to translate a string s, search for the sentence $\hat{t}$ that maximizes the probability that it was translated from s. Formally, this is modeled thus:

$$\hat{t} = \arg\max_t P(t \mid s) = \arg\max_t P(s \mid t) \times P(t)$$

To find $\hat{t}$, one has to estimate $P(t)$ and $P(s \mid t)$, and search for the t that maximizes their product. $P(t)$ is the *language model*; it is estimated from a monolingual corpus in the target language. $P(s \mid t)$ is the *translation model*; it is estimated from a *parallel corpus*, in which sentences in the source language are aligned with their translations in the target language. The search for an optimal translation is then implemented by a *decoder*. To improve the quality of the output of an SMT system, each of these components can be improved. In this work we address both the language model and the translation model.

Until recently, the ontological status of a text (as being original or translated) was not taken into account when building SMT systems. Several recent works, however, underscore the relevance of translationese for SMT. First, *language models* that are compiled from translated, rather than original, texts yield better SMT systems (Lembersky et al., 2011, 2012b). Second, *translation models* can be adapted to translationese, thereby improving the quality of the translation (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013).

Kurokawa et al. (2009) were the first to show that translationese matters for SMT. They defined two translation tasks, English-to-French and French-to-English, and used a parallel (English-French) corpus in which the translation direction of each text was indicated. They showed that for the English-to-French task, translation models compiled from English-translated-to-French texts were better than translation models compiled from texts translated in the reverse direction; and the same holds for the reverse translation task. These results were corroborated by Lembersky et al. (2012a, 2013), who further demonstrated that translation models can be adapted to translationese, thereby improving the quality of SMT even further.

Lembersky et al. (2011, 2012b) focused on the *language* model. They built several SMT systems, for several pairs of languages: French-English, German-English, Italian-

English and Dutch-English. For each language pair, they built two systems, one in which the language model was compiled from original English text, and another in which the language model was compiled from text translated to English from each one of the languages. They showed that language models complied from translated texts better fit the reference set, in term of perplexity. Moreover, SMT systems that were constructed from translationese-based language models perform much better than those constructed from original language models. In fact, an original corpus must be ten times larger in order to yield the same translation quality as a translated corpus.

To benefit from these results, however, one has to know whether the (monolingual and parallel) texts used for training SMT systems are original or translated; such meta-information is typically unavailable. However, due to the unique properties of translationese, the ontological status of texts as original or translated can be determined automatically using text-classification techniques. Several works address this task, and we survey some of them here.

Baroni and Bernardini (2006) proposed an approach to automatically identify translationese, distinguishing between original and translated Italian texts with a high level of accuracy. To eliminate a potentially rich source of content-based information, they replaced all words tagged as proper nouns by a special token. Each text was represented as a feature vector, where features include unigrams, bigrams and trigrams of various lexical pieces of data: the form, lemma, part of speech (POS) tag of words, as well as combinations thereof. They also created a *mixed* representation, where content words were replaced by their corresponding POS tags. As the value of each feature, they experimented with both the raw frequency and the *tf*idf* score. They used *recall maximization*, which labels an article as translated if at least one classifier predicts it as such, to combine the output of the classifiers. The results showed that most of the models outperform the random baseline (50%). To improve the results, they selected several sets of feature combinations, reaching over 85% accuracy for some combinations. Analysis of the results reveals that it is possible to identify translated texts merely based on function word distributions and shallow syntactic patterns.

Ilisei et al. (2010) checked whether markers of simplified language, such as lower type-to-token ratio, can contribute to classification accuracy. They employed different types of machine-learning classifiers (decision trees, naïve Bayes, SVM, etc.) with and without the

feature being inspected. The learning systems exploit twenty-one language-independent features, such as lexical richness and the proportion of grammatical words to lexical words, to decide if a text is original or translated. They found that the distribution of POS ungirams, augmented with lexical richness, proportion of grammatical words to lexical words, sentence length, and word length, yielded better results than using POS ungirams alone, thus corroborating previous non-computational studies in translation studies.

Koppel and Ordan (2011) used a list of 300 function words to construct a classifier for English; the classifier was tested on translations from other languages, as well as on texts in different genres. They used translations from five languages (Finnish, French, German, Italian and Spanish) to English, as well as original English, from the Europarl corpus (Koehn, 2005). For each language, they created a corpus consisting of 200 equal chunks of translated texts and 200 equal chunks of original English texts. Each chunk was represented as a vector of size 300 in which each entry was the frequency of the corresponding feature (a function word, taken from a list of most frequent English words) in the chunk. To understand how much of translationese is general, and how much of it is language dependent, they trained a cross validation classifier on one corpus (e.g., translations from French) and tested it on each one of the other corpora (e.g., translations from Italian). For any given source language, within each corpus, the resulting accuracy was 95–98.3%. For cross classification, however, when training on language L_1 and testing on language L_2, the accuracy deteriorated, and depended on the degree of similarity between L_1 and L_2. Likewise, training on one genre and testing on another yielded poorer results.

Popescu (2011), however, achieved better results on a cross-classification task using character 5-grams as features (note that most function words are included in this set of features). The training set consisted of 19th century novels translated from German to English vs. original American English; the test set included novels translated from French to English vs. original British English.

Volansky et al. (Forthcoming) focused on the features of translationese (in English) from a translation theory perspective. They defined several classifiers based on various linguistically-informed features, implementing in a computational way several hypotheses of Translation Studies. The classifiers were trained and tested on text chunks of approximately 2000 words, ending on a sentence boundary. Since we use a reimplementation of the same classifiers as our departure point in this work, we detail below the motivation be-

hind each of the features they define. While Volansky et al. (Forthcoming) only identified English translationese, we extend the experimentation also to French; we list below the few adaptations we introduced to their classifiers in order to support French. Note that we also explore classifiers that yielded relatively low results under the assumption that they may catch properties which are relevant for good machine translation despite poor performance on identifying translationese. We will return to this point when we discuss classifiers based on punctuation marks in Section 5.

Lexical Variety Original texts are expected to use richer language than translated texts. Consequently, their type-token ratio (TTR) is expected to be higher. Volansky et al. (Forthcoming) implement three different TTR measures, following Grieve (2007).

Mean word length If translated texts are indeed simpler than originals, their mean word length (in characters) can be expected to be lower.

Syllable ratio Similarly, translated texts are expected to have fewer syllables per word. This feature calculates the ratio of syllables per word, where a syllable is approximated by vowel-sequences that are delimited by consonants (or space) in each token.

Lexical density Translations are expected to have lower lexical density than original texts. This feature is implemented as the ratio of 'non-lexical' words (neither nouns, adjectives, adverbs nor verbs) to the number of tokens.

Mean sentence length The average length of sentences (in tokens), assuming that for simplicity, translated texts consist of shorter sentences. While this assumption was proved to be false, we keep this classifier as it performs above chance level (65% accuracy).

Mean word rank The assumption is that translated texts consist of more frequent words than original ones. This feature is implemented as the average rank of the words in each text, where the rank of a word is the position of the word in a pre-defined frequency-ordered list. The list is extracted from a very large corpus that models English more generally.

Most frequent words The frequencies of the N most frequent words in the corpus, where N is 5, 10, and 50. The assumption is that translated texts use frequent words more frequently than originals.

Explicit naming Translators show a tendency to use proper names as a clarification of personal pronouns. This feature computes the ratio of personal pronouns to proper nouns in the text.

Single naming and Mean multiple naming Translated texts are assumed to be more explicit, often elaborating when a person is mentioned. This feature calculates the frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor; and the average length (in tokens) of proper name sequences in each text.

Cohesive markers Cohesive markers enable readers to follow the rhetorical or discourse arrangement of the ideas in the text; translated texts are expected to use certain cohesive markers more frequently. This feature is implemented as the frequency of each cohesive marker (from a list of 40 markers) in the text.

Repetitions Translators tend not to repeat themselves. This feature counts sequences of content words that occur more than once in a chunk.

Contractions Translators tend to use more formal register, and therefore avoid contractions. This feature counts the ratio of contracted forms to their counterpart full form(s).

PMI Various theories predict different distributions of collocations across original and translated texts. Two measures test this hypothesis: the average *pointwise mutual information* (PMI) of all bi-grams in the text; and the count of bi-grams with PMI above 0.

Part-of-speech n-grams The source language is expected to leave traces of its syntactic structure on the translation product. To test this, the actual counts of every POS 1-, 2-, and 3-gram in each text is used as a feature.

Character n-grams The counts of character 1-, 2-, and 3-grams in each text. This feature captures lexical and morphological properties of the text.

Contextual function words Counts 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. This is a combination of lexical and shallow syntactic information.

Positional Token Frequency The choice of words at the begging and end of sentences is quite limited. This restriction can inflict differences between original and translated texts. This feature counts the first, second, ante-penultimate, penultimate, and last words of each sentence.

Function words Function words are known to be useful for various text classification tasks. This feature is implemented as the frequency of each function word, taken from a list of 467 words (Koppel and Ordan, 2011). For French classifiers, we use a similar list of French function words, listed in Appendix 5.

Pronouns This feature is inspired by Koppel and Ordan (2011); it is the frequency of each pronoun in the text.

Punctuation The counts of each punctuation mark in the text. Translated texts are expected to use punctuations differently, as they tend to be more explicit and less ambiguous.[1]

Ratio of passive forms Passive forms show higher incidence in English. Thus the use of passive form might be more common in original texts.

We reimplemented all these classifiers, and use the more accurate ones to predict translationese (see Section 4).

[1]For French classifiers, we added the $\ll$ and $\gg$ quotation marks to the list of punctuation marks used by Volansky et al. (Forthcoming).

3 Experimental setup

The experiments we describe in Section 4 consist of three parts:

1. Building classifiers to tease apart original from translated texts.

2. Constructing SMT systems with language models compiled from the predicted translations, comparing them with similar SMT systems whose language models consist of the entire monolingual corpora.

3. Constructing SMT systems with translation models compiled from bitexts that are predicted as translated in the same direction as the direction of the SMT task, comparing them with similar SMT systems whose translation models consist of the entire parallel corpora.

In this section we describe the language resources and tools required for performing these experiments.

3.1 Tools

Our first task is text classification; to ensure that the length of each text does not influence the classification, we partition the training corpus in most experiments into chunks of approximately 2000 tokens (ending on a sentence boundary). The notion of *chunk* is henceforth used to define the size of a sub-corpus. Our major experiments involve 2,500 chunks (of approximately 2,000 tokens each, hence 5M tokens). To detect sentence boundaries, we use the UIUC CCG sentence segmentation tool.[2]

We use the MOSES toolkit (Koehn et al., 2007) for tokenization and for converting the text to lower case. For English POS tagging we use *OpenNLP*[3] and for French POS tagging we use the *Stanford* tagger.[4] For classification we use *Weka* (Hall et al., 2009), a popular suite of machine learning tools. In all experiments we use the *SMO* algorithm, a support-vector machine with a linear kernel, in its default configuration.

To construct language models and measure perplexity, we use *SRILM* (Stolcke, 2002) with interpolated modified Kneser-Ney discounting (Chen, 1998) and with a fixed vocabulary. We limit language models to a fixed vocabulary and map out-of-vocabulary (OOV)

[2] `http://cogcomp.cs.illinois.edu/page/tools_view/2`, accessed 11.10.2013.
[3] `http://opennlp.apache.org`, accessed 24.08.2012.
[4] `http://nlp.stanford.edu/software/tagger.shtml`, accessed 08.02.2013.

tokens to a unique symbol to overcome sparsity and better control the OOV rates among various corpora.

We train and build the SMT systems using the MOSES toolkit (Koehn et al., 2007). For evaluation, we use MultEval (Clark et al., 2011), which takes machine translation hypotheses from several runs of an optimizer and provides three popular metric scores, BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and TER (Snover et al., 2006)), as well as standard deviations (via bootstrap resampling) and p-values (via approximate randomization).

3.2 Corpora

To construct SMT systems we need both monolingual corpora (for the language model) and bilingual ones (for the translation model). The main corpora we use are Europarl (Koehn, 2005) and the Canadian Hansard. The Europarl corpus is a multilingual corpus recording the proceedings of the European Parliament. Some portions of the corpus are annotated with the original language of the utterances, and we use the method of Lembersky et al. (2012a) to identify the source language of other sentences in this corpus. The Hansard corpus is a parallel corpus consisting of transcriptions of the Canadian parliament in English and (Canadian) French from 2001-2009. We use a version that is annotated with the original language of each parallel sentence. We also use the News Commentary corpus (Callison-Burch et al., 2007), a French-English parallel corpus in the domain of general politics, economics and science. The direction of translation of this corpus is not annotated.

We use the above mentioned corpora for training and for evaluation. We would have liked to have more diverse corpora for evaluation; unfortunately, we were unable to find other parallel, sentence-aligned corpora in which the direction of translation is indicated, and in which the original text units are retained. We hope that such corpora become available in the future, and we urge the developers of language resources to pay attention to the direction of translation and to retain, as much as possible, the original structure of the data so that coherent text chunks (in particular, consisting solely of original or solely or translated language) can be easily identified.

3.2.1 Language model experiments

This section describes the corpora we use for constructing language models; Section 3.2.2 describes the data we use for constructing translation models. Our main experiments focus on French translated to English (FR→EN), and we define classifiers that can identify English translationese. However, to further establish the robustness of our approach, we also experiment with German translated to English (DE→EN) and with English translated to French (EN→FR). We also conduct cross-corpus (X-corpus) experiments in which we train on one corpus and test on another.

From the Europarl corpus we use several portions, collected over the years 1996 to 1999 and 2001 to 2009. Table 1 lists some statistics (number of sentences, number of tokens and average sentence length) of the corpora used for training translationese classifiers and for training language models. For most experiments, the split of the monolingual corpora to translated vs. original texts is to equal portions (in terms of chunks), but we also experiment with unbalanced splits. The parallel corpora are divided to two sections according to the direction of the translation (when it is known). For example, for the French-to-English translation task, we divide the Europarl corpus to a French-original section (FR→EN) and an English-original section (O–EN).

Task	Corpus	Lang.	Portion (%)	Sentences	Tokens	Length
FR→EN	Europarl	FR→EN	50	85,750	2,546,085	29.7
	Europarl	O–EN	50	99,300	2,545,891	25.6
DE→EN	Europarl	DE→EN	50	87,900	2,322,973	26.4
	Europarl	O–EN	50	91,100	2,324,745	25.5
EN→FR	Europarl	EN→FR	50	95,378	2,761,334	28.9
	Europarl	O–FR	50	88,844	2,783,677	31.3
X-corpus, FR→EN	Hansard	FR→EN	50	241,044	4,002,648	16.6
	Hansard	O–EN	50	221,023	4,001,030	18.1
X-corpus, FR→EN	Hansard	FR→EN	25	44,919	799,993	17.8
	Hansard	O–EN	75	412,247	7,198,307	17.5
X-corpus	News			153,577	3,999,556	26.0

Table 1: Monolingual corpora used for training translationese classifiers and language models

We also use portions of the Europarl corpus to define reference sets with which we evaluate the perplexity of language models. For this task we only use translated texts. The sizes of these reference set are listed in Table 2.

For constructing translation models (to train SMT systems), we use parallel corpora. For the FR→EN and EN→FR tasks we use original French text (O–FR), aligned with its

Lang.	Sentences	Tokens
FR→EN	8,494	260,198
DE→EN	6,675	178,984
EN→FR	4,284	125,590

Table 2: Reference sets used for evaluating perplexity

translation to English (FR→EN). For the DE→EN translation task we use original German text (O–DE), aligned with its translation to English (DE→EN). The parallel portions we use are disjoint from those used for the language model and are evenly balanced between the original text and the aligned translated text. From Europarl we use portions from the period of January to September 2000. Table 3 lists statistics on the corpora used for training the translation models.

Task	Corpus	Lang.	Sentences	Tokens	Length
FR→EN, EN→FR	Europarl	FR→EN	88,996	2,312,798	26.0
	Europarl	O–FR	88,996	2,532,780	28.5
DE→EN	Europarl	DE→EN	89,810	2,389,418	26.6
	Europarl	O–DE	89,810	2,240,491	24.9
X-corpus, FR→EN	Hansard	FR→EN	683,264	12,188,277	17.8
	Hansard	O–FR	683,264	14,703,494	21.51

Table 3: Parallel corpora used for training translation models

To tune and evaluate SMT systems we use reference sets that are extracted from a parallel, aligned corpus. These include 1000 sentence pairs for tuning and 1000 (different) sentence pairs for evaluation. The sentences are randomly extracted from another portion of the Europarl corpus, collected over the period of October to December 2000, and another portion of Hansard; see Table 4. All references are of course disjoint from the training materials.

3.2.2 Translation model experiments

In this set of experiments we focus again on French translated to English (FR→EN) systems. However, to further establish the robustness of our approach, as in the language model experiments, we also experiment with German translated to English (DE→EN) and with English translated to French (EN→FR). We conduct four experiments: three in-domain experiments using the Europarl corpus, and one cross-corpus (X-corpus) experiments in which we train on one corpus and test on another.

From the Europarl corpus we use several portions, collected over the years 1996 to 1999

Task	Corpus	Lang.	Use	Sentences	Tokens	Length
FR→EN	Europarl	FR→EN	Tune	1000	31,335	31.3
	Europarl	O–FR	Tune	1000	32,182	32.1
	Europarl	FR→EN	Eval	1000	28,410	28.4
	Europarl	O–FR	Eval	1000	29,294	29.2
DE→EN	Europarl	DE→EN	Tune	1000	26,450	26.4
	Europarl	O–DE	Tune	1000	23,716	23.7
	Europarl	DE→EN	Eval	1000	25,608	25.6
	Europarl	O–DE	Eval	1000	23,184	23.2
EN→FR	Europarl	EN→FR	Tune	1000	30,111	30.1
	Europarl	O–EN	Tune	1000	25,940	25.9
	Europarl	EN→FR	Eval	1000	28,102	28.1
	Europarl	O–EN	Eval	1000	24,236	24.2
X-corpus, FR→EN	Hansard	FR→EN	Tune	1000	20,137	20.1
	Hansard	O–FR	Tune	1000	24,453	24.4
	Hansard	FR→EN	Eval	1000	17,631	17.6
	Hansard	O–FR	Eval	1000	21,492	21.4

Table 4: Reference sets used for evaluating SMT systems

and 2001 to 2009. Table 5 lists statistics on the corpora used for training translationese classifiers and for training translation models.

Task	Corpus	Lang.	Sentences	EN Tokens	FR/DE Tokens
FR→EN, EN→FR	Europarl	FR→EN	163,377	4,635,836	4,936,261
	Europarl	O–EN	131,677	3,292,458	3,796,657
DE→EN	Europarl	DE→EN	120,804	3,164,994	2,913,099
	Europarl	O–EN	127,363	3,191,971	3,189,011
X-corpus, FR→EN	Hansard	FR→EN	289,685	5,745,766	7,003,250
	Hansard	O–EN	294,420	5,797,516	6,901,689

Table 5: Parallel corpora used for training translationese classifiers and translation models

To construct language models, in the in-domain experiments we use Europarl portions from the period of January to September 2000 (this is the English /French side of the training data used for building the translation model in the corresponding language model experiments). For Hansard cross-corpus experiments, we use the language model built from translated texts that we use in the Hansard language model experiments. Table 6 lists statistics on the corpora used for the language models.

To tune and evaluate SMT systems we use the same reference sets used in the language model experiments; see Table 4. All references are disjoint from the training materials.

Task	Corpus	Lang.	Sentences	Tokens	Length
FR→EN	Europarl	FR→EN	88,996	2,312,798	26.0
DE→EN	Europarl	DE→EN	89,810	2,389,418	26.6
EN→FR	Europarl	O–FR	88,996	2,532,780	28.5
X-corpus, FR→EN	Hansard	FR→EN	241,044	4,002,648	16.6
X-corpus, FR→EN	Europarl		85,750	2,546,085	29.7

Table 6: Monolingual corpora used for language models

4 Experiments and results

4.1 Language models experiments

We focus on the language model in this section. We build different SMT systems that use the same translation model, but differ in their language models. We define the following tasks:

1. Train classifiers to detect translationese. This has been done before, and we re-implement the classifiers of Volansky et al. (Forthcoming), adding one additional classifier. We evaluate the accuracy of these classifiers intrinsically, using ten-fold cross-validation.

2. Build language models from the chunks that were predicted (by each classifier) as translated. We evaluate the accuracy of the classifiers extrinsically, computing the perplexity of those language models with respect to the reference sets (Table 2).

3. Construct SMT systems with these language models. Our hypothesis is that language models compiled from (predicted) translationese will perform as well as (or even better than) language models compiled from the entire corpus. We evaluate this hypothesis in several scenarios: when the corpus used for the language model is the same corpus used for training the classifiers; or a different one, but of the same type; or from a completely different domain.

We now detail the experiments and report the results.

4.1.1 Classification of translationese

We begin by re-implementing the 32 classifiers defined by Volansky et al. (Forthcoming) to tell translations from originals. We also implement one additional classifier, combining the feature sets of the contextual function words (CFW) and punctuations classifiers. To reproduce the results of Volansky et al. (Forthcoming), we use ten-fold cross-validation on the training corpus (Section 3.2) to evaluate the accuracy of these classifiers.

Then, we use the prediction of each classifier to determine whether test texts are original or translated. Thus, each classifier defines a partition of the training corpus to (predicted) originals vs. (predicted) translations. In other words, we treat each classifier output as a sub-corpus, consisting of all the text chunks that were classified as translated by this

classifier. Based only on each classifier's prediction, we build language models from the sub-corpora determined as translations. We then evaluate the fitness of these sub-corpora to the reference set, in terms of perplexity. Specifically, we train 1-, 2-, 3-, and 4-gram language models for each sub-corpus (a total of 32×4 language models) and measure their perplexity on the reference set.

				Perplexity			
	Features enabled	Chunks	Acc. (%)	1-gram	2-gram	3-gram	4-gram
1	Type/token ratio	1424	52.44	473.75	97.85	73.98	71.07
2	Ttr2	1591	52.64	473.57	96.70	72.54	69.59
3	Ttr3	1373	58.93	471.35	97.29	73.64	70.81
4	Mean word length	1259	63.39	468.93	96.93	73.66	70.85
5	Syllables ratio	1286	55.39	470.33	97.69	74.30	71.46
6	Lexical density	1274	67.09	468.47	97.23	73.71	70.79
7	Mean sentence length	1172	70.59	465.56	96.36	73.36	70.58
8	Mean word rank1	1243	70.59	471.96	98.16	74.76	71.98
9	Mean word rank2	1254	62.39	470.05	97.28	73.99	71.25
10	N top frequent words	1275	68.64	468.14	96.49	73.05	70.24
11	Positional token frequency	1265	**98.40**	**463.39**	**94.61**	**71.35**	**68.49**
12	Explicit naming	1749	52.88	472.17	95.73	**71.28**	**68.21**
13	Single naming	1799	59.96	470.24	**94.74**	**70.26**	**67.15**
14	Mean multiple naming	1090	60.24	468.82	98.68	75.66	72.95
15	Cohesive markers	1164	82.84	464.38	95.82	72.83	70.04
16	Pronouns	1324	78.63	465.20	95.05	**71.63**	**68.74**
17	POS-tag unigrams	1246	**95.50**	**463.40**	**94.73**	**71.54**	**68.68**
18	POS-tag bigrams	1255	**98.48**	**463.28**	**94.66**	**71.43**	**68.57**
19	POS-tag trigrams	1255	**99.44**	**463.39**	**94.72**	**71.47**	**68.60**
20	Letter unigrams	1248	77.27	465.42	95.83	72.51	69.69
21	Letter bigrams	1255	**97.93**	**463.50**	**94.82**	**71.56**	**68.69**
22	Letter trigrams	1257	**99.52**	**463.49**	**94.71**	**71.47**	**68.62**
23	Contextual function words	1255	**99.84**	**463.37**	**94.69**	**71.46**	**68.60**
24	Repetitions	1421	51.65	473.22	97.65	73.81	70.87
25	Contractions	42	50.93	495.15	187.48	174.15	173.32
26	Average PMI	1164	51.96	474.85	100.02	76.68	73.88
27	Threshold PMI	1010	54.59	471.25	101.39	78.35	75.63
28	Function words	1254	**96.77**	**463.59**	**94.79**	**71.53**	**68.69**
29	Punctuation	1199	**92.83**	**463.91**	95.30	72.17	69.33
30	PunctuationRatio1	1196	**92.79**	**463.79**	95.31	72.21	69.36
31	PunctuationRatio2	1212	**91.28**	464.10	95.28	72.11	69.26
32	Ratio of passive forms	1057	54.47	471.19	99.99	77.04	74.37
33	CFW and Punctuation	1245	**98.96**	**463.51**	**94.81**	**71.60**	**68.76**
34	Translated texts	1255		**463.58**	**94.59**	**71.24**	**68.37**
35	Original texts	1258		500.56	115.48	91.14	88.31
36	All texts	2513		473.00	93.34	**67.84**	**64.47**

Table 7: Classification of translationese, and fitness to the reference set of FR→EN language models compiled from texts predicted as translated

The results are reported in Table 7. Replicating the results of Volansky et al. (Forth-

coming), we demonstrate that some of the classifiers are indeed excellent, with accuracies of over 90%; the best-performing classifiers are highlighted in boldface in Table 7. Not surprisingly, the good classifiers yield better language models. The rightmost columns of Table 7 list the perplexity of language models trained on the sub-corpora that were predicted as translations, when applied to the reference set; several of the high-accuracy classifiers indeed induce language models whose perplexity is lower (hence better). For comparison, we provide in Table 7 also the perplexity of language models compiled from the entire training set (row 36); from the *actual* (as opposed to predicted) translated texts (row 34); and from the actual *original* texts (row 35). Clearly, and consistently with the results of Lembersky et al. (2012b), the original texts yield the worst language model (highest perplexity), whereas the actual translated texts yield an upper bound (lowest perplexity). Still, due to the high accuracy of some of the classifiers, their perplexity is very close to this upper bound. Row 36 reflects a model built from all texts, both original and translated; such a corpus is twice as large as the corpus used for the other models, hence the lower perplexity rates.

To further establish the robustness of these results, we repeat the experiments with other corpora, this time consisting of German translated to English (DE→EN), and also English translated to French (EN→FR). The latter experiment required an adaptation of some of the classifiers from English to French. We only focus on eleven of the best-performing classifiers in the new experiments; and we only report results for the 4-gram language models.

Table 8 presents the results of the DE→EN experiments. The accuracies of the classifiers are high, comparable to the case of FR→EN. Moreover, the perplexities of several of the induced language models are very close to the upper bound obtained by taking actual translated texts. Similarly, Table 9 reports the results of the EN→FR experiments, demonstrating a very similar pattern.

These results provide initial support to our hypothesis: in all three cases, language models that rely on the predictions of the better classifiers are almost as good as the language model compiled from (actual, rather than predicted) translated texts. We also measured the correlation between the accuracy of the classifiers and the perplexity rate. When only the highly accurate classifiers are taken into account (those whose accuracy is over 90%), this correlation is excellent, at 0.91. Evidently, accurate identification of translationese

	Feature enabled	Chunks	Acc. (%)	Perplexity
11	Positional token frequency	1,152	**97.35**	**62.18**
17	POS-tag unigrams	1,144	89.80	62.62
18	POS-tag bigrams	1,147	**97.83**	62.26
19	POS-tag trigrams	1,138	**98.56**	62.31
20	Letter unigrams	1,182	72.81	63.05
21	Letter bigrams	1,155	**96.35**	62.35
22	Letter trigrams	1,151	**99.21**	**62.16**
23	Contextual function words	1,151	**99.47**	**62.17**
28	Function words	1,154	**96.83**	62.27
29	Punctuation	1,118	84.43	63.17
30	PunctuationRatio1	1,126	84.43	63.15
31	PunctuationRatio2	1,112	81.65	63.56
33	CFW and Punctuation	1,146	**99.08**	62.23
34	Translated texts	1,153	-	**62.07**
35	Original texts	1,153	-	76.68
36	All	2,306	-	**57.48**

Table 8: Accuracy of the classification, and fitness of language models compiled from texts predicted as translated to the reference set, DE→EN

	Feature enabled	Chunks	Acc. (%)	Perplexity
11	Positional token frequency	1,421	**98.22**	47.91
17	POS-tag unigrams	1,433	**95.18**	**47.71**
18	POS-tag bigrams	1,411	**96.24**	**47.83**
19	POS-tag trigrams	1,413	**97.80**	**47.84**
20	Letter unigrams	1402	78.22	48.13
21	Letter bigrams	1,411	**97.09**	47.93
22	Letter trigrams	1,416	**99.53**	47.88
23	Contextual function words	1,418	**98.68**	47.91
28	Function words	1,418	**97.48**	**47.83**
29	Punctuation	1,435	**93.98**	**47.80**
30	PunctuationRatio1	1,434	**93.66**	**47.83**
31	PunctuationRatio2	1,433	**92.56**	**47.80**
33	CFW and Punctuation	1,410	**98.47**	47.92
34	Translated texts	1,413	-	47.89
35	Original texts	1,411	-	59.75
36	All	2,824	-	**44.49**

Table 9: Accuracy of the classification, and fitness of language models compiled from texts predicted as translated to the reference set, EN→FR

corresponds to better language models (in terms of perplexity).

4.1.2 Language models compiled from predicted translationese

Above, we established the fact that translated texts can be identified with high accuracy, and that language models compiled from predicted translations fit the reference sets well. In this section we explore the hypothesis that such language models are indeed good for

SMT.

We begin with a French-to-English translation task. We use the same language models described in Section 4.1.1, constructed from the predictions of the best performing classifiers. We also fix a single translation model, compiled from the parallel portion of the training corpus (Section 3.2). We then train several French-to-English SMT systems with each of the (predicted) LMs. All systems are tuned on the same tuning set of 1000 parallel sentences, and are tested on the same reference set of 1000 parallel sentences (Table 4). As a baseline, we build an SMT system that uses the entire training corpus for its language model; we refer to this system as *All*. As an upper bound, we build a system that uses the (actual) translated texts for its LM. We also report results on a system that uses only original texts for its LM.

We evaluate the quality of each of the SMT systems using MultEval (Section 3.1). The results are presented in Table 10, reporting the BLEU, METEOR, and TER evaluation measures, as well as the p-value defining the statistical significance with which the system is different from the baseline (with respect to the BLEU score only).

	Feature enabled	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	28.8	33.1	54.0	0.00
17	POS-tag unigrams	**29.0**	**33.3**	**53.7**	0.69
18	POS-tag bigrams	**29.0**	33.2	**53.7**	0.64
19	POS-tag trigrams	**28.9**	33.2	**53.8**	0.09
20	Letter unigrams	28.8	33.2	**53.9**	0.01
21	Letter bigrams	**29.1**	**33.3**	**53.7**	0.74
22	Letter trigrams	**28.9**	33.2	**53.8**	0.06
23	Contextual function words	**29.0**	**33.3**	**53.7**	0.67
28	Function words	**29.0**	**33.3**	**53.8**	0.42
29	Punctuation	28.8	33.1	**53.9**	0.02
30	PunctuationRatio1	**28.9**	**33.2**	**53.9**	0.09
31	PunctuationRatio2	**28.9**	**33.3**	**53.8**	0.14
33	CFW and Punctuation	**28.9**	**33.2**	**53.8**	0.16
34	Translated texts	**29.1**	**33.3**	**53.6**	0.58
35	Original texts	27.8	32.9	54.7	0.00
36	All	**29.1**	**33.3**	**53.8**	

Table 10: Evaluation of the FR→EN SMT systems built from language models compiled from predicted translationese

Replicating some of the results of Lembersky et al. (2011, 2012b), these results demonstrate that using only translated texts for the language model is not inferior to using the entire corpus (although the size of the latter is double the size of the former). In terms of BLEU scores, both yield the same score, 29.1. Similarly, as reported by Lembersky et al.

(2011, 2012b), using only original texts is markedly worse, with a BLEU score of 27.8. The main novelty of our current results, however, is the observation that several of the language models that only use *predicted*, rather than actual translated texts, perform just as well. In Table 10 we highlight in boldface entries that correspond to classifiers whose performance is better or not significantly different from the performance of the *All* classifier; there are several such classifiers, with BLEU scores between 28.9 and 29.1. The other two evaluation measures show the same pattern exactly.

For completeness, we repeat the same experiments with two more language pairs: German to English and English to French. The setup is identical, and we report the same evaluation metrics. The results are presented in Table 11 (German to English) and Table 12 (English to French). The emerging pattern is identical to the case of French to English.

	Feature enabled	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	21.6	28.5	63.9	0.02
17	POS-tag unigrams	21.6	28.4	64.5	0.00
18	POS-tag bigrams	**21.8**	**28.6**	64.0	0.31
19	POS-tag trigrams	**21.8**	**28.6**	**63.9**	0.35
20	Letter unigrams	**21.7**	28.4	64.2	0.12
21	Letter bigrams	21.6	28.5	64.2	0.01
22	Letter trigrams	**21.7**	**28.6**	64.1	0.13
23	Contextual function words	**21.8**	**28.6**	**63.9**	0.59
28	Function words	**21.7**	28.5	64.0	0.18
29	Punctuation	21.6	**28.7**	64.0	0.00
30	PunctuationRatio1	21.5	28.5	64.3	0.00
31	PunctuationRatio2	21.6	**28.6**	64.1	0.00
33	CFW and Punctuation	**21.9**	**28.6**	**63.8**	0.87
34	Translated texts	**21.8**	**28.6**	**63.9**	0.37
35	Original texts	21.0	28.4	64.6	0.00
36	All	**21.9**	**28.6**	**63.7**	-

Table 11: Evaluation of the DE→EN SMT systems built from language models compiled from predicted translationese

The results of all the experiments confirm our hypothesis; SMT systems built from *predicted* translationese language models perform as well as SMT systems built from (actual) translated language models, and similarly to (twice as large) mixed language models. The main finding of these experiments is that using predicted translations, rather then real translations, is indistinguishable from the upper bound. Several of our classifiers induce language models that yield BLEU scores of $28.9 - 29.1$ in FR→EN experiments, BLEU scores of $21.7 - 21.9$ in DE→EN experiments and BLEU scores of $26.2 - 26.3$ in EN→FR experiments, which are not significantly different from the upper bound and the baseline

	Feature enabled	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	**26.3**	47.8	**58.3**	0.54
17	POS-tag unigrams	**26.2**	**47.9**	**58.6**	0.20
18	POS-tag bigrams	**26.2**	47.7	**58.6**	0.06
19	POS-tag trigrams	**26.2**	**47.9**	**58.5**	0.22
20	Letter unigrams	25.9	47.6	59.2	0.00
21	Letter bigrams	26.1	47.8	**58.7**	0.03
22	Letter trigrams	26.1	47.8	**58.7**	0.02
23	Contextual function words	**26.3**	47.8	**58.3**	0.37
28	Function words	**26.3**	47.8	**58.9**	0.44
29	Punctuation	**26.3**	47.7	**58.5**	0.61
30	PunctuationRatio1	**26.2**	47.8	**58.8**	0.11
31	PunctuationRatio2	**26.2**	**47.9**	**58.8**	0.16
31	CFW and Punctuation	**26.3**	47.8	**58.3**	0.47
34	Translated texts	26.1	47.7	**58.5**	0.03
35	Original texts	25.1	47.0	59.5	0.00
36	All	**26.3**	**48.0**	**58.7**	-

Table 12: Evaluation of the EN→FR SMT systems built from language models compiled from predicted translationese

systems. Crucially, the size of the predicted language models is half the size of the baseline model.

4.1.3 Cross-corpus experiments

The experiments discussed above all use the same type of corpus both for training the translationese classifiers and for training the SMT systems (the actual portions differ, but all are taken from the same corpus). In a typical translation scenario, a monolingual corpus is available for constructing a language model, but the status of its texts (original or translated) is unknown, and has to be predicted by a classifier that was trained on a potentially different domain. The question we investigate in this section, then, is whether a classifier trained on texts in one domain is useful for predicting translationese in a different domain.

As a first experiment, we use (English) translationese classifiers that are trained on the Europarl training data, but use the Hansard training data for constructing the SMT system. In this experiment, we do not use the meta-information of the Hansard corpus, but instead use the predictions of the classifiers. Based on the prediction of each classifier, we define a partition of the Hansard training corpus to (predicted) originals vs. (predicted) translations and use the text chunks that were classified as translated by each classifier to build 4-grams language models.

Again, as in the in-domain experiment, we construct a single, fixed translation model

from the parallel portion of the (Hansard) corpus. We then train several French-to-English SMT systems with each of the (predicted) LMs. All systems are tuned and tested on the same tuning and evaluation reference set (Table 4). As a baseline, we build an SMT system that uses the entire Hansard training corpus for its language model; we refer to this system as *All*. As an upper bound, we build a system that uses the (real) translated texts for its LM. We also report results on a system that uses only original texts for its LM.

As we show in Table 13, the results are consistent with the findings of the in-domain experiments. The best-performing systems use either actual translated texts (BLEU score of 38.0), or the entire corpus (38.0); the worst system uses original texts (37.5, significantly below the baseline system). Several of the predicted-translationese systems perform at $37.8 - 38.0$, statistically insignificant difference compared with the upper bound. Note that some of the classifiers perform extremely poorly in this scenario, but still manage to identify chunks that are excellent for training language models; the punctuation classifiers are an example. We do not yet fully understand this phenomenon. Having said that, other classifiers, notably contextual function words, are both accurate and useful for SMT.

	Feature enabled	Chunks	Acc. (%)	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	601	60.21	37.1	37.3	46.5	0.00
17	POS-tag unigrams	1,045	49.87	37.6	**37.6**	46.2	0.01
18	POS-tag bigrams	1,226	65.64	37.4	37.5	46.1	0.00
19	POS-tag trigrams	872	67.29	**37.8**	37.6	**45.9**	0.11
20	Letter unigrams	2,296	49.90	**37.8**	**37.7**	**45.9**	0.03
21	Letter bigrams	1,936	64.39	37.7	**37.6**	**45.9**	0.03
22	Letter trigrams	616	62.54	37.5	37.5	46.1	0.00
23	Contextual function words	1,377	80.50	**37.8**	**37.7**	**45.9**	0.21
28	Function words	1,629	72.41	**37.9**	**37.7**	**45.8**	0.55
29	Punctuation	2,426	34.78	**37.9**	**37.7**	**45.9**	0.21
30	PunctuationRatio1	2,254	34.58	**38.0**	**37.7**	**45.8**	0.66
31	PunctuationRatio2	236	48.97	37.1	37.3	46.5	0.00
33	CFW and Punctuation	1,321	78.22	**37.8**	**37.7**	**45.9**	0.11
34	Translated texts	2001		**38.0**	**37.8**	**45.7**	0.86
35	Original texts	2001		37.5	37.6	46.1	0.00
36	All	4002		**38.0**	**37.7**	**45.8**	-

Table 13: Cross-corpus evaluation: Hansard-based SMT system, Europarl-based classification

The above experiments assume that the monolingual corpus is balanced: half of it is original and half of it is translated. While it is impossible to know what portion of a 'typical' corpus is translated (in some languages, mainly low-resource ones, the vast majority of texts may be translated (Pym and Chrupała, 2005)), we experiment also with a scenario in

which only a small portion (25%) of the corpus is translated, and the rest (75%) is original. The motivating question for this experiment is whether it is worthwhile to classify such a corpus and use for the language model only the (presumably very small) portion of texts predicted as translations.

Table 14 reports the results of this experiment. Observe that there is no statistically significant difference between the accuracy obtained with the entire corpus, only the one-fourth translated texts, and the three quarters of original texts. This demonstrates the trade-off between having fewer, but better (translated) texts and having more, but not as good (original) ones. Several of the classifiers are reasonably accurate in this scenario, identifying 1,000$\pm$250 chunks as translated, but they yield poor SMT systems. The only exceptions are the POS unigram and bigram classifiers, yielding BLEU scores that are not significantly worse than the ones obtained with the upper bounds.

Interestingly, some classifiers (letter unigrams and two punctuation classifiers) miserably fail to identify translations, but retain a large portion of the corpus, yielding excellent SMT systems. Especially with respect to the punctuation classifiers, we suspect that they may be able to identify patterns that end up useful for training language models, but much further analysis is required in order to better understand this.

	Feature enabled	Chunks	Acc. (%)	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	396	78.2	37.1	37.4	46.4	0.00
17	POS-tag unigrams	971	64.62	**37.7**	**37.6**	46.1	0.10
18	POS-tag bigrams	898	77.00	**37.7**	37.6	**45.8**	0.05
19	POS-tag trigrams	552	83.05	37.6	37.5	46.0	0.01
20	Letter unigrams	2,284	56.10	**37.9**	**37.7**	**45.8**	0.55
21	Letter bigrams	1,370	70.75	37.6	37.5	46.1	0.00
22	Letter trigrams	393	80.52	37.5	37.5	46.0	0.00
23	Contextual function words	746	89.35	37.5	37.6	46.1	0.00
28	Function words	834	83.05	37.5	37.5	46.1	0.00
29	Punctuation	2,802	29.77	**37.9**	**37.7**	**45.8**	0.57
30	PunctuationRatio1	2,630	31.67	**37.9**	**37.7**	**45.9**	0.81
31	PunctuationRatio2	244	72.85	37.2	37.4	46.3	0.00
33	CFW and Punctuation	740	88.1	37.6	37.5	46.1	0.01
34	Translated texts	1000	-	**37.8**	**37.7**	**45.9**	0.22
35	Original texts	3000	-	**37.7**	**37.6**	**45.9**	0.11
36	All	4000	-	**37.9**	**37.7**	**45.8**	-

Table 14: Cross-corpus evaluation, unbalanced split of O and T texts

We repeat the cross-corpus experiments with an additional corpus: the News Commentary corpus. This is a French-English parallel corpus, for which the direction of translation is not annotated; we only use its English side. Presumably, most of the texts in this corpus

consist of original English, but we hypothesize that the classifiers may be able to select chunks with translationese-like features and consequently provide better SMTs systems. Additionally, as the News Commentary corpus is a collection of editorials, we partition the corpus into (not necessarily equal-length) articles, rather than to 2000-token chunks. The motivation for this experiment is to check whether automatic prediction of translationese, applied to new texts from a new domain, and training language models on the predicted texts only, can yield SMT systems that perform as well as ones that use the entire corpus.

We thus train English translationese classifiers on both the Europarl and the Hansard monolingual corpora (Table 1), and use them to classify the News Commentary corpus. As in the first experiment, we use the prediction of the classifiers to define a partition of the News Commentary corpus to (predicted) originals vs. (predicted) translations and use the text chunks that were classified as translated by each classifier to build 4-gram language models. We then train several French-to-English SMT systems with each of the (predicted) LMs. All systems use the same Europarl translation model, the one used in the in-domain task (Tables 3, 4). As a baseline, we build an SMT system that uses the entire corpus for its language model. The results are presented in Table 15.

	Feature enabled	Chunks	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	1,931	**27.1**	**33.0**	**55.1**	0.40
17	POS-tag unigrams	2,511	**27.2**	**33.1**	**55.1**	0.95
18	POS-tag bigrams	1,782	**27.1**	**33.0**	**55.3**	0.16
19	POS-tag trigrams	1,756	27.0	**33.0**	**55.2**	0.02
20	Letter unigrams	1,622	**27.1**	**33.0**	**55.0**	0.41
21	Letter bigrams	2,233	**27.2**	**33.0**	**55.0**	0.65
22	Letter trigrams	1,959	**27.1**	**33.0**	**55.1**	0.40
23	Contextual function words	1,496	**27.1**	**33.0**	**55.0**	0.20
28	Function words	1,930	**27.2**	**33.1**	**55.0**	0.72
29	Punctuation	2,282	27.0	**33.0**	**55.2**	0.00
30	PunctuationRatio1	2,072	27.0	**33.0**	**55.0**	0.04
31	PunctuationRatio2	2,371	**27.2**	**33.1**	**55.1**	0.46
33	CFW and Punctuation	1,470	27.0	**33.0**	**55.2**	0.02
36	All	2,527	**27.2**	**33.0**	**55.2**	-

Table 15: Cross-corpus evaluation: LMs constructed from predicted translationese, News Commentary corpus

These experiments reveal the same pattern: several of the predicted-translationese systems yield BLEU scores of $27.1 - 27.2$, statistically insignificant difference compared with the All system that uses the entire corpus (27.2). This is obtained with much smaller corpora, e.g., only 1,930 chunks in the case of function words, or 1,496 chunks in the case of

contextual function words, compared with the entire corpus of 2,527 chunks.

4.2 Translation model experiments

We now move to experiments that address the translation model. We build SMT systems that use a fixed language model but differ in their translation model training data. For all systems we use fixed tuning and evaluation sets. We define the following tasks:

1. Train the same 33 classifiers to detect the direction of the translation (FR→EN vs. EN→FR). We classified the English side of the parallel corpora; for the FR→EN and DE→EN tasks, chunks predicted as translated are assumed to be translated in the right direction ($S \rightarrow T$). For the EN→FR task, chunks predicted as original are assumed to be translated in the right direction. We evaluate the accuracy of these classifiers intrinsically, using ten-fold cross-validation.

2. Build translation models from the English chunks that were predicted (by each classifier) as translated (original for EN→FR task), along with their aligned chunks in French.

3. Construct SMT systems with these translation models. Our hypothesis is that translation models compiled from parallel texts that are predicted as translated in the right direction ($S \rightarrow T$) will perform as well as (or even better than) translation models compiled from the entire corpus. We evaluate this hypothesis in two scenarios: when the corpus used for the translation model is the same corpus used for training the classifiers, and when the two corpora differ.

We now detail these experiments and report their results.

4.2.1 Translation models compiled from predicted translationese

We first train the 33 classifiers on the English side of the parallel corpora and use ten-fold cross-validation on the training corpus (Section 3.2) to evaluate the accuracy of these classifiers. Then, we use the prediction of each classifier to determine whether test texts are original or translated. Thus, each classifier defines a partition of the training corpus to (predicted) originals vs. (predicted) translations. For the FR→EN and DE→EN tasks, English translated texts are in the right translation direction ($S \rightarrow T$) and original texts are in the opposite direction ($T \rightarrow S$); for the EN→FR task English original texts are in the

right direction ($S \rightarrow T$) and translated texts are in the opposite direction ($T \rightarrow S$). We thus only use the chunks predicted as translations for the FR→EN and DE→EN tasks, and the chunks predicted as originals for the EN→FR task. For each partition, we match the English with the aligned French sentences, thereby defining the SMT training data.

We hypothesize that translation models built from such training data are better for SMT. To explore this hypothesis we fix a single language model (Section 3.2), and train several SMT systems with each of the (predicted) partitions and their aligned sentences. All systems are tuned on the same tuning set and are tested on the same reference set (Table 4). As a baseline, we build an SMT system, *All*, that uses the entire training corpus for its translation model. As an upper bound, we build a system that uses for its translation model the portion of the parallel corpus that was indeed translated in the right direction ($S \rightarrow T$). We also report results on a system that uses only the portion of the parallel corpus that was translated in the opposite direction ($T \rightarrow S$) for its translation model. The results are presented in Table 16.

	Feature enabled	Chunks	Acc. (%)	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	1,686	98.16	**31.2**	**34.7**	**52.1**	0.16
17	POS-tag unigrams	1,679	95.05	31.0	34.6	52.2	0.00
18	POS-tag bigrams	1,692	97.63	30.9	34.6	52.2	0.00
19	POS-tag trigrams	1,682	98.93	**31.1**	**34.8**	**52.0**	0.20
20	Letter unigrams	1,718	79.11	30.8	34.4	52.6	0.00
21	Letter bigrams	1,705	97.42	31.0	**34.7**	52.2	0.00
22	Letter trigrams	1,688	99.28	**31.2**	**34.8**	**52.0**	0.18
23	Contextual function words	1,682	99.46	**31.2**	**34.8**	**52.1**	0.23
28	Function words	1,693	96.77	**31.2**	**34.7**	**51.9**	0.34
29	Punctuation	1,613	90.20	30.9	34.4	52.5	0.00
30	PunctuationRatio1	1,635	88.55	30.8	34.4	52.5	0.00
31	PunctuationRatio2	1,628	90.23	30.6	34.3	52.4	0.00
33	CFW and Punctuation	1,678	98.93	**31.1**	**34.7**	**52.1**	0.13
34	$S \rightarrow T$	1,690	-	**31.3**	**34.8**	**51.7**	0.94
35	$T \rightarrow S$	1,689	-	28.4	33.3	54.4	0.00
36	All	3,379	-	**31.3**	**34.7**	**51.9**	-

Table 16: Accuracy of the classification and evaluation of the FR→EN SMT systems built from translation models compiled from predicted translationese

Again, the results demonstrate that some of the classifiers are indeed excellent, with accuracies of over 90%. These results are consistent with previous works that showed that SMT systems trained on $S \rightarrow T$ parallel texts outperformed systems trained on $T \rightarrow S$ texts (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013). Indeed, the best-performing systems use either (actual) $S \rightarrow T$ texts (BLEU score of 31.3), or the entire corpus (31.3);

the worst system uses (actual) $T \rightarrow S$ texts (28.4). What we add to previous results is the corroboration of the hypothesis that predicted-translationese systems perform just as well as the actual ones: indeed, several of the classifiers yield BLEU scores of $31.1 - 31.2$, a statistically insignificant difference compared with the upper bound.

As in the language model experiments, we repeat the same experiments with two more language pairs: German to English and English to French. The setup is identical, and we report the same evaluation metrics. The results are presented in Table 17 (German to English) and Table 18 (English to French).

	Feature enabled	Chunks	Acc. (%)	BLEU↑	METEOR↑	TER↓	p-value
11	Positional Token frequency	1,618	98.35	23.9	30.3	61.5	0.00
17	POS-tag unigrams	1,598	89.61	23.7	30.1	61.5	0.00
18	POS-tag bigrams	1,608	97.55	23.9	30.2	61.6	0.01
19	POS-tag trigrams	1,605	99.00	23.6	30.2	61.9	0.0
20	Letter unigrams	1,639	85.12	23.6	30.1	61.5	0.00
21	Letter bigrams	1,607	1,618	23.7	30.2	61.7	0.00
22	Letter trigrams	1,611	99.56	23.7	30.3	61.8	0.00
23	Contextual function words	1,608	99.78	23.8	30.0	61.6	0.00
28	Function Words	1,624	97.3	23.8	30.3	61.6	0.00
29	Punctuation	1,544	81.18	23.7	30.1	61.8	0.00
30	PunctuationRatio1	1,536	80.99	23.7	30.1	61.8	0.00
31	PunctuationRatio2	1,549	80.90	23.7	30.1	61.8	0.00
33	CFW and Punctuation	1,607	99.44	23.7	30.3	61.6	0.00
34	$S \rightarrow T$	1,613	-	**24.0**	30.4	**61.3**	0.05
35	$T \rightarrow S$	1,612	-	21.7	29.0	63.9	0.00
36	All	3,225	-	**24.2**	**30.5**	**61.1**	-

Table 17: Accuracy of the classification and evaluation of the DE→EN SMT systems built from translation models compiled from predicted translationese

The emerging pattern is similar to the case of French to English, confirming our hypothesis: SMT systems built from *predicted* $S \rightarrow T$ systems perform as well as SMT systems built from the entire corpus. In the case of DE→EN (Table 17), two classifiers induce SMT systems whose performance is close to the one built from real translated texts (although using the entire corpus is still significantly better). In the case of EN→FR, however, Table 18 shows that some classifiers induce SMT systems that are even better than ones compiled from real translations, and significantly better than systems built from the entire corpus.

4.2.2 Cross-corpus experiments

In the experiments discussed in the previous section we used the same corpus both for training the translationese classifiers and for training the SMT systems (the actual portions

	Feature enabled	Chunks	Acc. (%)	BLEU↑	METEOR↑	TER↓	p-value
11	Positional Token frequency	1,693	98.16	**29.4**	**50.7**	**56.2**	0.04
17	POS-tag unigrams	1,700	95.05	**29.3**	**50.7**	**55.4**	0.23
18	POS-tag bigrams	1,687	97.63	**29.0**	**50.7**	56.5	0.36
19	POS-tag trigrams	1,697	98.93	**29.3**	**50.8**	**56.2**	0.19
20	Letter unigrams	1,661	79.11	**29.1**	**50.6**	56.4	0.79
21	Letter bigrams	1,674	97.42	**29.3**	**50.9**	**56.0**	0.18
22	Letter trigrams	1,691	99.28	**29.3**	**50.8**	**56.2**	0.16
23	Contextual function words	1,697	99.46	**29.3**	**50.8**	**56.0**	0.16
28	Function Words	1,686	96.77	**29.2**	**50.6**	**55.8**	0.73
29	Punctuation	1,766	90.20	**29.2**	**50.8**	**56.0**	0.72
30	PunctuationRatio1	1,744	88.55	**29.4**	**50.8**	**56.2**	0.05
31	PunctuationRatio2	1,751	90.23	**29.1**	**50.6**	**56.1**	0.51
33	CFW and Punctuation	1,678	98.93	**29.4**	**50.7**	**55.3**	0.11
34	$S \rightarrow T$	1,689	-	**29.3**	**50.8**	**56.1**	0.18
35	$T \rightarrow S$	1,690	-	26.7	48.2	58.2	0.00
36	All	3,379	-	**29.1**	**50.6**	**56.0**	-

Table 18: Accuracy of the classification and evaluation of the EN→FR SMT systems built from translation models compiled from predicted translationese

differ, but all are taken from the same corpus). The question we investigate in this section is whether a classifier trained on texts in one domain is useful for predicting translationese in a different domain.

We use (English) translationese classifiers that are trained on the Europarl training data, but use the Hansard corpus for the translation model. We apply the Europarl-trained classifiers to the English side of the Hansard corpus, and based on the prediction of each classifier, define a partition of the Hansard training corpus to use for the translation model. As in the in-domain experiment, we construct a single, fixed language model from a portion of the (Hansard) corpus. We then train several French-to-English SMT systems with each of the (predicted) translation models. All systems are tuned and tested on the same tuning and evaluation reference set (Table 4). As a baseline, we build an SMT system that uses the entire Hansard training corpus for its translation model; we refer to this system as *All*. As an upper bound, we build a system that uses the (actual) $S \rightarrow T$ texts for its translation model. We also report results on a system that uses only $T \rightarrow S$ texts for its translation model.

Table 19 reports the results of this experiment. The best-performing systems use either actual $S \rightarrow T$ texts or the entire corpus (BLEU score of 37.3). Only few classifiers (letter unigrams, letter bigrams and two punctuation classifiers) retain a portion of the corpus that is large enough for yielding viable SMT systems. Three of these classifiers perform at

	Feature enabled	Chunks	Acc. (%)	BLEU↑	METEOR↑	TER↓	p-value
11	Positional token frequency	497	56.75	32.5	35.0	49.8	0.00
17	POS-tag unigrams	1,933	72.41	36.5	36.9	46.8	0.00
18	POS-tag bigrams	766	61.43	33.3	35.6	49.0	0.00
19	POS-tag trigrams	1,189	69.61	35.2	36.6	47.6	0.00
20	Letter unigrams	3,179	74.61	36.9	**37.3**	46.3	0.01
21	Letter bigrams	2,373	73.8	36.4	36.8	46.9	0.00
22	Letter trigrams	688	60.66	34.2	35.8	48.3	0.00
23	Contextual function words	1,779	78.75	36.3	36.9	46.5	0.00
28	Function words	1,610	71.16	36.3	36.9	47.0	0.00
29	Punctuation	2,496	80.46	36.9	37.1	46.7	0.02
30	PunctuationRatio1	2,387	79.41	**37.1**	**37.3**	46.4	0.21
31	PunctuationRatio2	569	57.68	32.3	35.1	50.0	0.00
33	CFW and Punctuation	1,840	79.36	36.3	36.9	46.6	0.00
34	$S \rightarrow T$	3,000	-	**37.3**	**37.3**	**46.2**	0.94
35	$T \rightarrow S$	3,000	-	34.1	35.8	48.9	0.00
36	All	6,000	-	**37.3**	**37.4**	**46.0**	-

Table 19: Cross-corpus evaluation: Hansard-based SMT system, Europarl-based classification

$36.9 - 37.1$, a small difference compared with the upper bound, (although the size of the latter is bigger). When METEOR scores are considered, these three classifiers yield scores that are not significantly different from the upper bound.

5 Discussion

Awareness to translationese is useful for machine translation. Numerous works have established the fact that translated texts are markedly different from original ones, to the extent that trained classifiers can identify translationese with high accuracy, in some scenarios reaching 100% (Baroni and Bernardini, 2006; Ilisei et al., 2010; Koppel and Ordan, 2011; Avner, 2013; Volansky et al., Forthcoming). Another insight was that translations from different source languages are also different from each other and can be well predicted. The third point about translationese is typological: translations from typologically closer source languages are closer to each other (Koppel and Ordan, 2011).

This work relies on two additional insights (Kurokawa et al., 2009; Lembersky et al., 2012b, 2013):

1. Direction matters. When constructing translation models from parallel texts it is important to identify which side of the bitext is the source and which is the target. Translation from the source of the SMT task to its target is always better than the reverse option.

2. Translationese matters. When constructing language models, translated texts (especially from the source language, but not only) are preferable to texts written originally in the target language of the task at hand.

Specifically, we train classifiers to identify translationese, and then use their predictions to construct language- and translation-models for SMT, demonstrating that attention to translationese can yield state-of-the-art translation quality with only a fraction of the corpora.

More and more data are available over the Internet these days, including parallel and monolingual texts which under the current statistical paradigm are central to machine translation. Available data are often noisy, and even state-of-the-art commercial translation machines are troubled by the effect noisy data has on translation quality (Venugopal et al., 2011). To overcome this problem, one has to better differentiate between different kinds of data.

For most works in SMT, corpora are black boxes: parallel and monolingual data in the target language are required, and no assumptions and therefore no preferences are made regarding these data. It has been shown, however, that conditioning the language model on a text's meta-information, such as the identity of the speaker or the session itself (in the

Hansard corpus), can improve machine translation (Foster et al., 2010). Like Foster et al. (2010), we find that considering sentences as the basic units of consideration is at present naïve: classifiers of translationese require larger chunks of texts to perform at a reasonable level. In future work we would like to improve our translationese classifiers such that smaller chunks of text suffice for accurate identification of translationese. Meanwhile, developers of language resources should strive to retain the original structure of collected texts, ideally keeping paragraph boundaries.

Like Volansky et al. (Forthcoming), we find that the best performing classifiers are those that reflect *interference*, that is, reflect traces left on the target text from the source language. These features do not lend themselves easily to analysis because in order to understand them better one has to sort weighted features and consider them against the source text; such analysis is beyond the scope of this work.

As a rule of thumb, one can rely on classifiers that identify at least half of the data as translated for both the language model and the translation model. Having said that, note that in some cases even a third of the data sufficed to achieve results which were as good as using the entire corpus (function words and contextual function words and, Table 13).

Throughout all the experiments, the punctuation-based classifiers yielded good SMT systems even in cases where they miss-identified many original texts as translated. This may be due to the fact that they spot certain features that are good for statistical machine translation but are not necessarily translations themselves. In future work we intend to further explore this phenomenon and look beyond translationese towards texts which are translationese-like at least in the sense that they optimize translation scores.

Some of the classifiers yield SMT systems that perform at quality not significantly different from the quality of systems compiled from known translationese. With only one exception, all the classifiers we used implement exactly one set of translationese features; we believe that combined classifiers, implementing more than one set of features, may further improve the classification accuracy and therefore the SMT quality. We only experiment with one such combination (contextual function words and punctuation); more work is required to identify the best feature combination for this task.

Finally, we mainly experimented with English and French in this work, but we are confident that many language pairs can benefit from the methodology we propose. More work is needed to establish this empirically.

References

Ehud Alexander Avner. Identifying Hebrew translationese using machine learning techniques. Diplomarbeit, University of Potsdam, 2013.

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.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990. ISSN 0891-2017.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-0718.

Stanley F. Chen. An empirical study of smoothing techniques for language modeling. Technical report 10-98, Computer Science Group, Harvard University, November 1998.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-2031.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91. Association for Computational Linguistics, July 2011. URL http://www.aclweb.org/anthology/W11-2107.

George Foster, Pierre Isabelle, and Roland Kuhn. Translating structured documents. In *Proceedings of AMTA*, 2010.

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.

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.

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.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-2045.

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.

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.

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.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39, January 2013. URL http://dx.doi.org/10.1162/COLI_a_00159.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

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 and Grzegorz Chrupała. The quantitative analysis of translation flows in the age of an international language. In Albert Branchadell and Lovell M. West, editors, *Less Translated Languages*, pages 27–38. John Benjamins, Amsterdam, 2005.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006. URL http://www.cs.umd.edu/~snover/tercom/.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *Procedings of International Conference on Spoken Language Processing*, pages 901–904, 2002. URL citeseer.ist.psu.edu/stolcke02srilm.html.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1126.

Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Literary and Linguistic Computing*, Forthcoming.

Appendix A: French function words

Afin de, Afin de que, Ailleurs, Ainsi, Ainsi que, Alors, Alors que, Arrière, Assez, Auparavant, Aussitôt, Aussitôt que, Autour, Autour de, Autrefois, Autrement, Bien que, Car, Cependant, Certes, Contre, D'abord, D'ailleurs, Davantage, De plus, Debout, Dedans, Dehors, Depuis, Depuis que, Dessous, Dessus, Donc, Dès, Dès que, Désormais, En outre, Encore, Enfin, Ensuite, Envers, Environ, Hors, Hors de, Jadis, Jusqu'à ce queJusque, a cause de, a moins que, a peine, afin de, afin de que, afin que, ai, aie, aient, aies, ailleurs, ainsi, ainsi que, ait, alors, alors que, arrière, as, assez, au, au fur et a mesure, aucuns, auparavant, aura, aurai, auraient, aurais, aurait, auras, aurez, auriez, aurions, aurons, auront, aussi, aussi bien que, aussitôt, aussitôt que, autour, autour de, autre, autrefois, autrement, aux, avaient, avais, avait, avant, avec, avez, aviez, avions, avoir, avons, ayant, ayez, ayons, bien que, bon, c', car, ce, ceci, cela, celle -là, celle-ci, celles-là, celui-ci, celui-là, celà, cependant, certes, ces, cet, cette, ceux, ceux-là, chacun, chacune, chaque, ci, combien, comme, comme si, comment, contre, d', d'abord, d'ailleurs, dans, davantage, de, de plus, debout, dedans, dehors, depuis, depuis que, des, dessous, dessus, deux, devrait, doit, donc, dos, droite, du, dès, dès que, début, désormais, elle, elles, en, en même temps, en outre, encore, enfin, ensuite, envers, environ, es, essai, est, et, eu, eue, eues, eurent, eus, eusse, eussent, eusses, eussiez, eussions, eut, eux, eûmes, eût, eûtes, fait, faites, fois, font, force, furent, fus, fusse, fussent, fusses, fussiez, fussions, fut, fûmes, fût, fûtes, grâce à, haut, hors, hors de, ici, il, ils, j', jadis, je, jusqu'à ce que, jusque, juste, l', la, la leur, la mienne, la miennes, la nôtre, la sienne, la siennes, la tienne, la tiennes, laquelle, laquelles, le, le leur, le mien, le miens, le nôtre, le sien, le siens, le tien, le tiens, lequel, lequels, les, les leurs, leur, leurs, lors, lorsque, lui, là, m', ma, maintenant, mais, malgré, me, mes, mine, moi, moins, mon, mot, même, même si, n', n'importe qui, n'importe quoi, ne, ni, nommés, nos, notre, nous, nouveaux, néanmoins, on, ont, or, ou, où, par, par conséquent, par contre, parce, parfois, parmi, parole, partout, pas, pendant combien de temps, pendant que, personne, personnes, peu, peut, pièce, plupart, plutôt, pour, pour que, pourquoi, pourtant, pourvu que, presque, puis, puisque, qu, qu'est-ce que, quand, quand même, quant à, que, quel, quelle, quelles, quelqu'un, quelque chose, quelquefois, quelques-uns, quels, qui, quoique, rien, rien rien, s', sa, sans, sans que, sauf, se, sera, serai, seraient, serais, serait, seras, serez, seriez, serions, serons, seront, ses, seul, seulement, si, si bien que, sien, soi, soient, sois, soit, sommes, son, sont, sous, soyez, soyons, suis, sujet, sur, surtout, t', ta,

tandis, tandis que, tant que, te, tellement, tels, tes, toi, ton, toujours, tous, tous les deux, tout, tout de suite, tout le monde, toutefois, toutes, toutes les deu, trop, très, tu, tôt, un, une, valeur, vers, voie, voient, volontiers, vont, vos, votre, vous, vu, y, À moins que, À peine , à, à condition que, à quelle distance, à qui, ça, étaient, étais, était, étant, état, étiez, étions, été, étée, étées, étés, êtes, être