Following the trail of source languages in literary translations

Carmen Klaussner, Gerard Lynch and Carl Vogel

Abstract We build on past research in distinguishing English translations from originally English text, and in guessing the source language where the text is deemed to be a translation. We replicate an extant method in relation to both a reconstruction of the original data set and a fresh data set compiled on an analogous basis. We extend this with an analysis of the features that emerge from the combined data set. Finally, we report on an inverse use of the method, not as guessing the source language of a translated text, but as a tool in quality estimation, marking a text as requiring inspection if it is guessed to be a translation, rather than a text composed originally in the language analysed. With obtain c. 80% accuracy, comparable to results of earlier work in literary source language guessing—this supports the claim of the method's validity in identifying salient features of source language interference.

1 Introduction

The present study replicates and extends research reported on by Lynch and Vogel (2012), which sought to identify salient features in texts of translated English that discriminate them from texts originally composed in English. That study focused on an English literary text corpus, containing translations into English from French, German and Russian, as well as non-translated English, and achieved c. 80% accuracy. The feature set included 'document level' statistics, in addition to token-level distinctive features. By comparison, van Halteren (2008) achieved 86% accuracy in a similar task with speeches in the EUROPARL corpus, using token-level features.

Carmen Klaussner
CNGL, Trinity College Dublin, Ireland e-mail: klaussnc@tcd.ie

Gerard Lynch
CeADAR, University College Dublin, Ireland e-mail: gerard.lynch@ucd.ie

Carl Vogel
Centre for Computing and Language Studies, Trinity College Dublin, Ireland e-mail: vogel@tcd.ie

The task in the literary discrimination problem considered by Lynch and Vogel (2012) was to correctly identify the source language (L1) of a text, be it a translation or a text composed originally in English. They explored the classification efficacy of different feature types — document-level statistics, such as average word length or readability metrics, aggregated over the whole text and other features are individuated as n-grams of tokens used in the text, or abstracted as part-of-speech (POS) sequences.[1] The features identified can be regarded as tangible influences of the source language on the target language texts. Text selection was restricted to those originating in the late 19th to early 20th century as made available through the Project Gutenberg archive. To limit the effect of individual author (or translator) style, each author and translator was only represented once in the corpus, the aim being the detection of features of source-language translation effects rather than *fingerprints* of individual authors or translators. In the present replication of the task, we test the method's overall validity. For this purpose our data set is independent with respect to internal overlaps of translators and authors with distinct texts from those used in earlier work. We deem demonstration of replicability and extensibility to be an important part of scientific endeavour, and therefore we report our findings here.

Thus, section 2 gives an insight into more intrinsic causes of translation effects and recent work that studied their manifestations in translations. Section 3 describes the current data set and methodology for the experiment. In section 4, we present our experiments and results on the detection of source languages. Finally, section 5 closes with an in-depth analysis of the prediction properties of identified highly discriminatory attributes and compares to those identified on the basis of the previous study's data set. For the purpose of comparison, we further combine both data sets and subsequently analyse highly characteristic features. As a final exercise in quality estimation, we use salient features to see whether they will lead us to texts that contain irregularities in terms of *normal* English style.

2 Motivation & Previous Research

The task of predicting the source language of a text by identifying elements that indirectly originated in the source language presupposes *translation effects*, in this case influence of the source language on the target language. Since translators traditionally translate into their native language, an interesting question to consider is how differences between translated and original language might arise. The act of translation requires fidelity to the source text while maintaining adequacy of representation in the target text (Volansky et al., 2013), and given these two forces pulling in opposite directions, it might be expected that one of these factors will prevail.

Translation studies brought forth the idea of *translation universals*, properties inherent to the act of translation, such as *explicitation* (the tendency to render content that is expressed implicitly in the source language more explicitly the target lan-

[1] Lexical items that could identify works by topic or theme were discarded to retain and use only features widely applicable.

guage) or *simplification* (the act of simplifying content in the form of, for instance a less complex sentence structure) that were originally introduced by Blum-Kulka (1986). However, the concept of translation universals remains a controversial issue, partly connected to the difficulty of proving it and thus is constantly challenged by emerging research, as recently by Becher (2010) and by Volansky et al. (2013).

The property of *interference*, which is by nature a phenomenon specific to language pairs, is simpler to investigate, as irregularities in the target language (L2) can generally be traced back to properties of the source language (L1). Most of these influences are likely to be more abstract, and thus difficult to prevent, as well as being highly dependent on the particular translation pair and therefore not necessarily pertaining to universal translation effects. For instance, a study into bidirectional transfer has shown that one is influenced by a language's conceptual properties regardless of whether it was learned as a first or second language (Degani et al., 2011).[2]

Another issue is related to register interference for pairs of languages that have a similar syntactic realisation of certain items, but which do not share the same register or code relation, where two items in L2 of different register point to one lexical and register representation in L1. For example, in German *"Ich bin"* can translate to both *"I am"* and *"I'm"* in English. Whereas the German non-contracted version can also appear in colloquial speech, in contrast *"I am"* appears mostly in written text and would be rather marked elsewhere. However, a translator would need to be able to make this distinction (in an abstract way) and decide based on the context rather than only the lexical realisation and similarity to the target language construction. Practically, this is likely to result in more instances of *"I am"* in the target language. Also, interference may arise due to lack of language competence of the translator, for instance with respect to comprehending metaphors or idiomatic expressions to a degree that would allow rephrasing them appropriately given the cultural context of the target language. Failure of this sort may result in word-by-word translations that are structurally and lexically unusual in the target language.

Text classification methods trying to identify specific items of translated text likely originating in the source language have only recently been adopted. Work by Baroni and Bernardini (2006) spearheaded the usage of machine learning techniques to separate translated from non-translated text, in this case based on a large Italian newspaper corpus comprising both original texts and translations. A variety of feature types were explored, the most distinguishing being a combination of lemma and mixed unigrams, bigrams and POS trigrams resulting in ~86% classification accuracy. It showed that mandatory linguistic features in the source language that are optional in the target language are more likely to be overrepresented in a translation from L1 to L2. In work on the EUROPARL corpus, van Halteren (2008) was able to achieve 87.2% to 96.7% accuracy in guessing the source language of speeches, taking into account the signalling strength of individual sequences being present for texts originating in particular languages as well as degrees of over- or under-use. A number of studies explored translation direction and *translationese*,

[2] Consequently, both L1 and L2 speakers of Hebrew and English rated word pairs in English closer when they shared a translation in Hebrew, e.g. *tool* and *dish* both translate into *kli*. English monolinguals did not perceive this relatedness of the same word pair.

focused on machine translation. Kurokawa et al. (2009) investigated translation direction in the parallel Canadian Hansard FR-EN corpora and discovered that using FR-EN corpora for a FR-EN machine translation system resulted in the same accuracy (BLEU score) as a much larger corpus of EN-FR. Lembersky et al. (2011) extended this using both the EUROPARL and the Canadian Hansard corpus, focusing in the first instance on the language model used in machine translation, not the phrase tables themselves as done by Kurokawa et al. (2009). Using the *perplexity* metric on the EUROPARL data, they find that a language model made up of language translated from the source language of the text to be translated had the highest perplexity score, with the worst perplexity score derived from the original English data. A mixed model containing text from all four languages plus original English performed second best. They claim this validates the *convergence* universal of translation: a mixed *translationese* model was still a better fit than original English; *translationese* is more similar to other *translationese* from different source languages than original English. Lembersky et al. (2012) focused on the phrase tables themselves (like Kurokawa et al. (2009)), replicating the earlier results, delving further into the explanation of why translation direction matters. Using the information-theoretic metrics of *entropy* and *cross entropy*, they find that phrase tables trained in the correct direction contain more unique source phrases and less translations per source phrase, which results in a lower entropy and cross entropy score on the data.

Koppel and Ordan (2011) investigated dialects of *translationese*, which concerned the different subsets of translations in a target language from numerous source languages. Their work used the frequencies of the top three hundred most frequent words and SVM classifiers and obtained 90% classification accuracy between translations into English from Hebrew, Korean and Greek from the International Herald Tribune (IHT). They replicated and extended the work on EUROPARL from van Halteren (2008) and reported that results on corpora of same-language translations and original English are generally better than training a *translationese* detection system on one corpus (EUROPARL) and testing on another (IHT), which returned poor results (50-60%), showing that features of *translationese* vary not only by source language, but also genre. Ilisei et al. (2010) examined translation universals of simplification using document-level metrics only, and reported high accuracy (80-90%) on two separate corpora of medical and technical translations from Spanish to English. Distinguishing features included lexical richness, the ratio of closed-class to open-class words and the ratio of numerals, adjective and finite verbs, each to total words. This work motivated the inclusion of ‘document-level’ features in the original study on source language detection by Lynch and Vogel (2012), which is continued here. Similar methods are used in Lynch (2014) for the task of distinguishing authorial style in a corpus of literary translations of three Russian authors by the same translator, distinguishing reliably between authors using document-statistics alone. Forsyth and Lam (2014) also investigates authorial style in parallel translations of Van Gogh’s letters using clustering techniques and finds that authorial discriminability is stronger than translatorial discriminability.

Volansky et al. (2013) tested translation hypotheses, such as *simplification*, *explicitation* and *interference*, using supervised machine learning in English with linguistically informed features in a text classification setting. They tentatively conclude that claims to universal translation effects should be reconsidered, as they not only rely on genre and register, but also vary greatly depending on language pairings. *Interference* yields the highest performing and most straightforwardly interpretable distinguishing features ("source language shining through"). In the present study, we attempt to identify salient features of interference that originated in the source language, but are not translator-specific issues; i.e. features detected should be general enough to appear across different translations of the same source language and thus signal more intrinsic transfer issues for that particular language pair.

3 Methods

First, we replicated the study of Lynch and Vogel (2012), finding results that agree with the earlier work.[3] Below we describe our methods in extending this method to a new set of literary translations. We used the same restrictions introduced by the earlier study. The data selection process and preprocessing are described in section 3.1; feature sets, in section 3.2; classification methods, in section 3.3.

3.1 Corpus Selection & Preparation

In order to have a comparable data sample with respect to the first analysis, the present set was selected according to the same criteria of being available in machine-readable format and being in the public domain.[4] Each author should have a unique text and no translators should be repeated in the case of translations. Additionally, all texts should be of sufficient length, meaning at least 200 kilobytes in size.

Since this is the second experiment using this configuration, to be able to generalise better about the model's validity, there should be little or if possible no overlap in authors/translators with the first data set. While being able to meet all other requirements, the last one proved to be more difficult with respect to the Russian corpus, as there were fewer texts available (not theoretically, but practically, since new books are still in the process of being added) and after discarding all those not suited for other reasons, there remained one overlap with the first data set for the translator, *Constance Garnett*. Generally, one of the objectives was also to not have too large a time-frame, i.e. have all novels written and translated within 100 years, so as to limit differences in style with respect to diachronic effects. It would also have been more desirable for the corpus to contain only novels, whereas in the current set, there is one collection of short stories in the Russian set, namely *"The wife*

[3] See previous and replicated results in table 8 and 9 in the online appendix to this paper: https://www.scss.tcd.ie/~vogel/SGAI2014/KlaussnerLynchVogel-Appendix-SGAI2014.pdf

[4] All were taken from *Project Gutenberg*: http://www.gutenberg.org/ – last verified Nov. 2013.

Table 1 Text corpora for current source language detection experiment.

Title	Author	Source	Pub.	Translator	T. Pub.
Vanity Fair	William M. Thackeray	English	1848	n/a	n/a
Wives and Daughters	Elizabeth Gaskell	English	1866	n/a	n/a
The Moonstone	Wilkie Collins	English	1868	n/a	n/a
Sylvie and Bruno	Lewis Carroll	English	1889	n/a	n/a
The Return of Sherlock Holmes	Arthur Conan Doyle	English	1904	n/a	n/a
House of Mirth	Edith Wharton	English	1905	n/a	n/a
Only a girl: or, A Physician for the Soul	Wilhelmine von Hillern	German	1869	A. L. Wister	1870
Villa Eden	Berthold Auerbach	German	1869	Charles C. Shackford	1871
The Northern Light	E. Werner	German	1890	D.M. Lowrey	1891
Cleopatra - Complete	Georg Ebers	German	1893	Mary J. Safford	1894
Dame Care	Hermann Sudermann	German	1887	Bertha Overbeck	1891
Royal Highness	Thomas Mann	German	1909	A. Cecil Curtis	1916
The King of the Mountains	Edmond About	French	1856	Mrs. C. A. Kingsbury	1897
Against The Grain	Joris-Karl Huysmans	French	1884	John Howard	1924
The Dream	Emile Zola	French	1888	Eliza E. Chase	1893
Pierre and Jean	Guy de Maupassant	French	1888	Clara Bell	1890
Arsne Lupin versus Herlock Sholmes	Maurice Leblanc	French	1908	George Moorehead	1910
The Gods are Athirst	Anatole France	French	1912	Mrs. Wilfrid Jackson	1913
Dead Souls	Nicholas Gogol	Russian	1840	D. J. Hogarth	1846
The Wife, and Other Stories	Anton Chekhov	Russian	1892	Constance Garnett	1918
The Daughter of the Commandant	Aleksandr S. Pushkin	Russian	1836	Mrs. Milne Home	1891
The Precipice	Ivan Goncharov	Russian	1869	unknown/ M. Bryant	1916
A Russian Gentleman	Sergei T. Aksakov	Russian	1858	J.D. Duff	1917
Satan's Diary	Leonid Andreyev	Russian	1919	Herman Bernstein	1920

and other stories" rather than a continuous text. The difference in genre may affect the distribution of proper nouns, since there are likely to be more *different* ones.

In the original study, there were five novels per source language, while for ours, one more novel per source language corpus is added, as shown in table 1. We retain the same file sizes and partition scheme, whereby 200 kilobytes of text are randomly extracted from each novel and divided into pieces of 10 kilobytes each, resulting in 120 text chunks per source language. Limited availability of texts translated from the chosen languages, as well as our constraints on novel size and non-overlap with the previous experiment rendered a significantly larger data set infeasible.

3.2 Feature Sets

We consider the same set of document-level features as Lynch and Vogel (2012), shown in table 2, the inclusion of which had originally been motivated by the exploration of Ilisei et al. (2010) into features other than n-grams. In addition, we selected both part-of-speech (POS) trigrams (including all punctuation) and word unigrams. The previous study used POS bigrams but not POS trigrams. Another alteration in our system is the weighting scheme. In the previous study lexical and POS features were binarized: a feature was observed as either present or not regardless of the exact frequency magnitude. We adopt relative frequency weighting, which in this case is reduced to taking raw frequency counts, since all text chunks have about the same size and there is no need to normalise over document length. Binarized weighting was abandoned because of a considerable drop in performance on the current data set, while tests using relative frequency were successful on both data sets.

Table 2 Document level features

No	Feature	Description
1	Avgsent	Average sentence length
2	Avgwordlength	Average word length
3	ARI	Readability metric
4	CLI	Readability metric
No	**Feature**	**Ratio Description**
5	Typetoken	word types : total words
6	Numratio	numerals : total words
7	Fverbratio	finite verbs : total words
8	Prepratio	prepositions : total words
9	Conjratio	conjunctions : total words
10	Infoload	open-class words : total words
11	dmarkratio	discourse markers : total words
12	Nounratio	nouns : total words
13	Grammlex	open-class words : closed-class words
14	simplecomplex	simple sentences : complex sentences
15	Pnounratio	pronouns : total words
16	lexrichness	lemmas : total words
17	simpletotal	simple sentences : total sentences
18	complextotal	complex sentences : total sentences

Table 3 Top 20 ranked features selected for run 8-13.

Top 20 features	χ^2
it's	53.13
NP NP POS	52.35
IN NP NP	50.12
, VBN IN	46.79
, DT NN	43.26
, DT JJ	43.68
don't	43.56
got	43.46
DT NNS IN	43.40
and	42.63
NNS , DT	42.31
VBP JJ PP	41.70
suppose	40.38
NNS IN DT	40.35
i'll	40.11
IN NN ,	39.40
NN , VBN	39.23
NN , DT	38.79
Typetoken	38.77
NP NP SENT	38.54

3.3 Classification & Selection Methods

For processing of all features, we use R scripts dealing separately with word unigrams, POS n-grams and document-level features. POS tagging of the text is done using the R **koRpus** package for POS tagging (that uses TreeTagger POS tagger (Michalke, 2013; Schmid, 1994)). For both POS trigrams and word unigrams, we only retain the most frequent ones to dispose of highly novel-specific proper nouns and terms, as well as thus ensuring that we build a very general classifier.

The process of feature selection and classification remains the same as in the previous study. In order to preselect characteristic features from both the word unigram and POS trigram list, we rank each group separately according to classification power using the χ^2 metric, where a higher value indicates a greater divergence over different classes. We retain the most discriminatory 100 POS trigrams and 50 word unigrams. All 18 document-level features, shown in table 2, enter the final selection.

For machine-learning, we also used Weka (Hall et al., 2009), specifically the Naive Bayes classifier, the SMO classifier, which is an implementation of a Support Vector Machine (SVM) classifier and the Simple Logistic classifier.

4 Experiments

4.1 Method

For benefit of comparison, experiments are repeated using the same general settings as in the previous study by testing document-level features on their own and having two different configurations for selecting the highest discriminatory features from the preselected features (out of the two n-gram feature groups). In the 1st setting, we classify on only the 18 document-level statistics. For the 2nd setting, we select the best 50 features out of the preselected 100 best POS trigrams, 50 best unigrams and all document-level features, and for the 3rd configuration, rather than 50, we retain only the best 30 features according to χ^2 for the final model. The training and test sets are constructed by randomly extracting ~20% from the full set, resulting in a division of 390 instances for training and 90 instances for evaluation randomly spread over all four classes, which increases the proportion of test items in the total set by 3% compared to what was used previously. Using only the training data, we execute ten-fold cross validation; we apply separate evaluation against the test set.

4.2 Results

Table 4 shows the results for all experiments, using both test set and cross-validation for evaluation. In relation to the first study, results for document-level features in run 2-7 are occasionally both higher and lower. Similarly for the different configurations in runs 8-13 and runs 14-19, results move in the same range between ~60% to ~85% accuracy. Our highest value, also run 13 is slightly lower, but differences in terms of data set and training and test division probably account for that. Table 3 shows the first 20 features selected for run 8-13 (the full list is shown in table 11 of the online appendix – see fn. 3). In comparison to the earlier study, it is notable that the χ^2 value for the most discriminatory features is with 53.13 less than three times as low as the best in the previous study (χ^2 of 191.1184). Another general difference is the composition of the best 50 features. Although the same number of each feature type was considered, our set has a higher number of POS features and fewer unigrams (13), whereas in the previous study, 32 features out of 50 were word unigrams. This difference might be due to our discarding rather infrequent unigrams and thus not considering the complete set of features for preselection. However, it could also be due to POS trigrams (including punctuation) being better discriminators than POS bigrams, at least on the current data set. For a clearer analysis, we need to examine the features identified on the previous data set to determine consistent features over both data sets and therefore the ones that are more prevalent for a particular class.

Table 4 Results for model evaluation based on three different sets of features.

Run	Training	Test	Classifier	Features	Accuracy
1	Full	10-f-cv	Baseline	n/a	25.0%
2	Full	Test	NaiveBayes	18 doc-level	51.8%
3	Full	Test	SVM (SMO)	18 doc-level	63.0%
4	Full	Test	SimpLog	18 doc-level	67.9%
5	Full	10-f-cv	NaiveBayes	18 doc-level	56.2%
6	Full	10-f-cv	SVM (SMO)	18 doc-level	60.2%
7	Full	10-f-cv	SimpLog	18 doc-level	65.3%
8	Full	Test	NaiveBayes	Top50(100POStri+18doc+50wuni)	73.8%
9	Full	Test	SVM (SMO)	Top50(100POStri+18doc+50wuni)	79.7%
10	Full	Test	SimpLog	Top50(100POStri+18doc+50wuni)	76.9%
11	Full	10-f-cv	NaiveBayes	Top50(100POStri+18doc+50wuni)	77.2%
12	Full	10-f-cv	SVM (SMO)	Top50(100POStri+18doc+50wuni)	82.7%
13	Full	10-f-cv	SimpLog	Top50(100POStri+18doc+50wuni)	84.5%
14	Full	Test	NaiveBayes	Top30(100POStri+18doc+50wuni)	59.4%
15	Full	Test	SVM (SMO)	Top30(100POStri+18doc+50wuni)	76.2%
16	Full	Test	SimpLog	Top30(100POStri+18doc+50wuni)	71.1%
17	Full	10-f-cv	NaiveBayes	Top30(100POStri+18doc+50wuni)	69.7%
18	Full	10-f-cv	SVM (SMO)	Top30(100POStri+18doc+50wuni)	77.7%
19	Full	10-f-cv	SimpLog	Top30(100POStri+18doc+50wuni)	74.5%

5 Analysing Discriminatory Features

In this section, we explore salient features of this study, and the features obtained when we combine the two data sets, with respect to what they predict for the source languages. We close with an exercise in quality estimation using salient features.

5.1 Feature Prediction

To determine what features predict in terms of the four classes, we analyse those prominent 50 features that emerged from this study as being best in discriminating between classes. If a feature is frequent in a particular translation, but not in the English corpus, this might signal potential source language interference. We again

Table 5 Feature proportions over entire source language corpus. Bold features are those significantly more frequent and Italic features and those significantly less frequent than expected with respect to the feature type distribution over that source language.

Ranked Features	% English	% French	% German	% Russian
it’s	**61**	*3*	*19*	*17*
NP NP POS	**65**	*9*	*17*	*9*
IN NP NP	**48**	*17*	*16*	*18*
, VBN IN	*15*	**46**	*19*	*20*
, DT NN	*17*	**37**	*22*	24
, DT JJ	*16*	**40**	*21*	23
don’t	**42**	*9*	*16*	**33**
got	**48**	*10*	*17*	25
DT NNS IN	*19*	**34**	25	*22*
and	*23*	*23*	*25*	**29**
NNS , DT	*13*	**48**	*17*	23
VBP JJ PP	**64**	*6*	16	14
suppose	**59**	*10*	*15*	16
NNS IN DT	*21*	**35**	23	*21*
i’ll	**58**	*0*	27	*14*
IN NN ,	*20*	**29**	24	**27**
NN , VBN	*15*	**46**	*19*	*20*
NN , DT	*14*	**40**	*22*	24

extracted a sample and analysed each feature in all four source languages by taking the raw counts of the feature in the language and normalising it by the total count for its feature type in the language.[5] Table 5 shows the proportion of use for each source language in relation to the other languages for 18 of the 50 best features (table 10 in the online appendix contains the complete list—see fn. 3).[6] Most interesting are those where translations deviate vastly from the English source corpus, e.g. IT’S and I’LL—most uses of these items are in the English sample (61% and 58%, respectively) and least are in the French source corpus (3% and 0%, resp.). Differences in syntactic sequences can be seen for ⟨NP NP POS⟩,[7] a frequent sequence in English which appears less frequently in translations, or ⟨NNS , DT⟩, which is less common in originally composed English than in translations, especially from French.

5.1.1 Comparison Between Data Sets

To compare the two data sets (“A” for previous data and “B” for current) on equal terms, we consider the features emerging from only our system. Shared items among the top 50 features selected from each are indicated in table 6. That they are the most robust features over both sets indicates that they have strongest predictive power. However, there might be differences in what exactly they are predicting between

[5] A POS trigram, e.g. ⟨NP NP NP⟩, is relativized to the total number of witnessed POS trigrams.

[6] All proportions are rounded to the nearest integer.

[7] In this context, ⟨POS⟩ refers to “possessive ending” rather than part-of-speech (POS).

the data sets, and thus we need to look the distributions of the individual features in each set separately. Table 6 shows the raw counts for each feature and the total count of observed feature type (unigrams/POS trigrams). For each feature, the table also shows the percentage of its occurrences within texts of each language.[8] Consider ⟨DT NNS IN⟩ and ⟨, DT NN⟩, both relatively consistent in use between data sets. The use in English, German and Russian as source language is lower, and in French, higher than would expected if the feature did not depend on source language (with significant Pearson residuals, $p < 0.001$). Thus, these features are potentially reliable *interference* signals. However, other features have distributions that differ significantly between data sets. Consider IT’S, for example: the proportion in B for English is much greater than in A (61% vs. 28%), where the “Russian” sample has 57% of the instances of IT’S. A χ^2 comparison of corresponding tables of counts shows significant difference ($p < 0.001$), with significant Pearson residuals indicating that the frequency of IT’S for English in set B and Russian in set A lies well over expected counts for each. Even though this feature is discriminatory for both sets, in A it is indicative of translation from Russian and in B is diagnostic of English original texts. Thus, finding discriminatory features may be easier than finding true signals of source language interference. That candidate interference signals are not reliably diagnostic across data sets is revealed by replication studies like this.

Table 6 Comparison between prediction direction of shared features of both data sets by looking at proportion in % and raw counts # of shared items in both sets separately. The left and right table contain proportions of data set A and B respectively. Bold and Italic are used as in Table 5.

	Data set A								**Data set B**							
Features	Eng		Fr		Ger		Ru		Eng		Fr		Ger		Ru	
	%	(#)	%	(#)	%	(#)	%	(#)	%	(#)	%	(#)	%	(#)	%	(#)
it’s	28	(82)	*6*	(16)	*9*	(27)	**57**	(163)	**61**	(218)	*3*	(10)	*19*	(68)	*17*	(59)
don’t	**43**	(386)	*5*	(46)	23	(205)	**30**	(270)	**42**	(286)	*9*	(58)	*16*	(109)	33	(221)
, DT NN	*20*	(302)	**38**	(556)	*22*	(330)	*20*	(291)	*17*	(320)	**37**	(697)	*22*	(402)	24	(441)
DT NNS IN	*23*	(432)	**35**	(663)	*22*	(422)	*20*	(368)	*19*	(437)	**34**	(780)	25	(563)	22	(515)
which	*21*	(582)	**36**	(1011)	*24*	(680)	*18*	(505)	**24**	(1051)	**30**	(1286)	**28**	(1222)	*18*	(771)
NP NP NP	21	(163)	**50**	(383)	*19*	(144)	*11*	(82)	**39**	(280)	24	(171)	29	(206)	*9*	(62)
that’s	**33**	(96)	*5*	(14)	*11*	(31)	**52**	(153)	**48**	(105)	*3*	(7)	32	(71)	*17*	(38)
the	*23*	(9655)	**30**	(12262)	*25*	(10419)	*22*	(9176)	24	(12296)	**27**	(13869)	**26**	(13131)	*22*	(11286)
and	**27**	(6363)	*20*	(4866)	**27**	(6473)	**26**	(6092)	*23*	(6526)	*23*	(6466)	*25*	(6944)	**29**	(8179)
NN : SENT	*12*	(6)	*2*	(1)	*0*	(0)	**85**	(41)	*3*	(5)	**29**	(51)	16	(28)	**53**	(93)
Type totals																
unigram	147912		146380		150362		147637		174428		172429		177486		173300	
POS trigram	228923		223340		230062		226778		268539		260359		262624		264828	

5.1.2 Combined Data Set

As a final exercise, we combine both data sets into one, by discarding one novel per class in data set B in order to create a balanced set with respect to the two different

[8] This is relevant, since there are four more data samples in set B.

Table 7 18 most discriminatory features in the combined data set.

Sequence	χ^2	Data	Sequence	χ^2	Data	Sequence	χ^2	Data
NP NP POS	41.27	B	and	34.55	S	NN , DT	33.39	B
don’t	40.69	S	SENT NN VBZ	34.07	A	CD NNS ,	32.89	A
IN NP NP	38.11	B	SENT ” NN	33.90	A	SENT NNS VBP	32.07	A
, VBN IN	38.01	B	suppose	33.90	B	Conjratio	31.98	S
SENT CC PP	36.30	A	Nounratio	33.67	A	Avgwordlength	31.93	A
, DT NN	34.92	S	which	33.43	S	though	31.90	A

sets, i.e. have 5 novels per class per data set, thus 400 instances for each data set and 800 in total. We apply the settings from the most successful runs/configurations, 11-13, by preselecting 168 features out of 50 word unigrams, 100 POS trigrams and 18 document-level features and then again selecting the best 50 out of those using χ^2. The results although slightly lower than what was achieved earlier are still well in the range of the previous one (c. 75% to 82%). In table 7 the “Data” column indicates if the feature was prevalent in data set **A**, **B** or **S** if it was shared.

5.2 *An Exercise in Quality Estimation*

We have identified features in translated texts likely to have originated in the source language. However, since these were selected on the basis of a statistical optimisation, it is not guaranteed that texts containing these features would be judged equally anomalous by native speakers of English. Here, we examine whether highly salient features of our model index sentences that a human would strongly perceive as either original English or translated English. We selected a number of salient features for each source language and sentences in which they occur from the data set.

The gold standard notes whether an item was a translated text or originally composed in English. We obtained two human judgements on 17 sentences selected across our corpus.[9] The sentences were selected manually to maximise features distinctive for the source languages. The kappa statistic computed between the raters’ judgements is 0.406, suggesting moderate agreement. On 7 out of 17 instances, both raters agreed with the gold standard and on another 5, at least one was correct.[10]

Some sentences were identified by both annotators as translations, as for instance example (1), which has feature type ⟨, CC RB⟩ of *German* English. There is some disagreement between other ratings of translated pieces, but interestingly, not only the translated sentences introduced confusion: example (2) is an instance of original English (tagged to include ⟨IN NP NP⟩), but was judged translated by both raters.

[9] The assessors are native English speakers among co-authors of this paper.

[10] On 5 items, both raters agreed with each other but not with the gold standard: in 3 of those the source was English, and both raters judged the item odd; in the other 2, the source was not English.

(1) ... whereupon the Professor [...] had then expressed his thanks with glad surprise**, and indeed** emotion, for the kind interest the Prince had expressed in his lectures.

(2) The young couple had a house **near Berkeley Square** and a small villa at Roehampton, among the banking colony there.

This evidence constitutes anecdotal support for the claim that the method is correctly discriminative: items that are high in distinctive features that would be identified by the system slip by human scrutiny in 29% of cases, but agree with human judgement in 70% of cases. While this requires more elaborate testing and rigorous methodology (Schütze, 1996; Schütze, 2005), the results tentatively indicate that highly discriminatory features could be useful in quality estimation of translations.

6 Conclusion and Future Directions

We have replicated and extended past work in identifying linguistic features that discriminate the source language of literary texts presented in English, perhaps through translation. Our replication of work by Lynch and Vogel (2012), using the same source data, method and feature types,[11] yielded results which were a slight improvement on the original regarding the task of source language identification. Our reproduction of the experimental design on a fresh data set confirmed some signals of source language *interference* and showed others to be unreliable between data sets. This speaks to the need for still more extensive studies of this sort. We conclude that this method of source language detection on unseen texts of English from the literary genre is robust, and has potential for use in quality assessment for English prose. This work contributes to the emergent practice within the digital humanities of applying automatic text classification techniques to explore hypotheses.

Acknowledgements This research is conducted partly within the Centre for Global Intelligent Content (www.cngl.ie), funded by Science Foundation Ireland (12/CE/I2267).

1. Baroni, M. and Bernardini, S. “A new approach to the study of translationese: Machine-learning the difference between original and translated text”. In: *Literary and Linguistic Computing* 21.3 (2006), pp. 259–274.
2. Becher, V. “Abandoning the notion of ‘translation-inherent’ explicitation. Against a dogma of translation studies”. In: *Across Languages and Cultures* 11 (2010), pp. 1–28.
3. Blum-Kulka, S. “Shifts of cohesion and coherence in translation”. In: *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* 17 (1986), p. 35.
4. Degani, T., Prior, A., and Tokowicz, N. “Bidirectional transfer: The effect of sharing a translation”. In: *J. of Cog. Psychology* 23.1 (2011), pp. 18–28.

[11] Our current study employed POS trigrams, the original work used POS bigrams but not trigrams.

5. Forsyth, R. S. and Lam, P. W. “Found in translation: To what extent is authorial discriminability preserved by translators?” In: *Literary and Linguistic Computing* 29.2 (2014), pp. 199–217.
6. Hall, M. et al. “The WEKA data mining software: an update”. In: *ACM SIGKDD explorations newsletter* 11.1 (2009), pp. 10–18.
7. Ilisei, I. et al. “Identification of translationese: A machine learning approach”. In: *Computational Linguistics and Intelligent Text Processing*. Springer, 2010, pp. 503–511.
8. Koppel, M. and Ordan, N. “Translationese and Its Dialects”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. 2011, pp. 1318–1326.
9. Kurokawa, D., Goutte, C., and Isabelle, P. “Automatic Detection of Translated Text and its Impact on Machine Translation”. In: *Proceedings of the XII MT Summit,Ottawa, Ontario, Canada*. AMTA. 2009.
10. Lembersky, G., Ordan, N., and Wintner, S. “Language Models for Machine Translation: Original vs. Translated Texts”. In: *Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing*. 2011, pp. 363–374.
11. Lembersky, G., Ordan, N., and Wintner, S. “Adapting translation models to translationese improves SMT”. In: *Proceedings of the 13th Conference of the European Chapter of the Assoc. for Computational Linguistics, Avignon, France*. 2012, p. 255.
12. Lynch, G. “A Supervised Learning Approach Towards Profiling the Preservation of Authorial Style in Literary Translations”. In: *Proc. 25th COLING*. 2014, pp. 376–386.
13. Lynch, G. and Vogel, C. “Towards the Automatic Detection of the Source Language of a Literary Translation”. In: *Proc. 24th COLING (Posters)*. 2012, pp. 775–784.
14. Michalke, M. *koRpus: An R Package for Text Analysis*. (Version 0.04-40), last verified: 11.08.2014. 2013. URL: http://reaktanz.de/?c=hacking&s=koRpus.
15. Schmid, H. “Probabilistic part-of-speech tagging using decision trees”. In: *Proceedings of international conference on new methods in language processing*. Vol. 12. Manchester, UK. 1994, pp. 44–49.
16. Schütze, C. *The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology*. University of Chicago Press, 1996.
17. Schütze, C. “Thinking About What We Are Asking Speakers to Do”. In: *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Ed. by Kepsar, S. and Reis, M. Mouton De Gruyter, 2005, pp. 457–485.
18. van Halteren, H. “Source Language Markers in EUROPARL Translations”. In: *Proc. 22nd COLING*. 2008, pp. 937–944.
19. Volansky, V., Ordan, N., and Wintner, S. “On the features of translationese”. In: *Literary and Linguistic Computing* (2013).