A Supervised Learning Approach Towards Profiling the Preservation of Authorial Style in Literary Translations

Gerard Lynch
Centre for Applied Data Analytics Research
University College Dublin
Ireland
firstname.lastname@ucd.ie

Abstract

Recently there has been growing interest in the application of approaches from the text classification literature to fine-grained problems of textual stylometry. This paper seeks to answer a question which has concerned the translation studies community: how does a literary translator's style vary across their translations of different authors? This study focuses on the works of Constance Garnett, one of the most prolific English-language translators of Russian literature, and uses supervised learning approaches to analyse her translations of three well-known Russian authors, Ivan Turgenev, Fyodor Dosteyevsky and Anton Chekhov. This analysis seeks to identify common linguistic patterns which hold for all of the translations from the same author. Based on the experimental results, it is ascertained that both document-level metrics and n-gram features prove useful for distinguishing between authorial contributions in our translation corpus and their individual efficacy increases further when these two feature types are combined, resulting in classification accuracy of greater than 90 % on the task of predicting the original author of a textual segment using a Support Vector Machine classifier. The ratio of nouns and pronouns to total tokens are identified as distinguishing features in the document metrics space, along with occurrences of common adverbs and reporting verbs from the collection of n-gram features.

1 Introduction

The application of *supervised learning* technologies to textual data from the humanities in order to shed light on stylometric questions has become more popular of late. In particular, these approaches have been applied to questions from the field of translation studies, which concern the notion of *translationese*[1] detection in Italian and other languages, (Baroni and Bernardini, 2006; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Popescu, 2011; Koppel and Ordan, 2011; Lembersky et al., 2011). Work has also been carried out on source language detection from translation corpora, (van Halteren, 2008; Lynch and Vogel, 2012) and translation direction detection in parallel MT training corpora, (Kurokawa et al., 2009), which can have applications in the domain of machine translation where the direction of bilingual translation corpora has been shown to impact on the accuracy of automated translations using such corpora[2].

This work seeks to apply these methods to the task of identifying authorial style within a corpus of translations by the same translator. Venuti (1995) mentions the concept of the *translator's invisibility*, that the measure of the best translator is that their style is not distinguishable in the translation, that their main concern and focus is to deliver the original text in a faithful manner. Of course, this task is often subject to their own vocabulary choices and as was often the case, cultural or personal bias of the translator or the regime or government in which they were operating. Identifying the former case will

[1] The subset or dialect of language which consists solely of translations from another language.

[2] Translating FR-EN, a smaller bilingual corpus of French translated to English provides similar qualitative results (BLEU score) to a larger corpus consisting of English translated to French.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 376–386, Dublin, Ireland, August 23-29 2014.

be the focus of this work, as choices of vocabulary or sentence construction can be isolated through the application of machine learning methods, although the latter is also a highly interesting question, albeit a more complex one to tackle using the methods at hand.[3]

2 Previous work

Baroni and Bernardini (2006) were among the first to apply advanced machine learning techniques to questions of textual stylometry, although use of linguistic features and metrics was already established in studies such as Borin and Pruetz (2001) who worked on POS distributions in translated Swedish and work by Mikhailov and Villikka (2001) who examined translated Finnish using statistical methods and metrics from authorship attribution. Baroni and Bernardini (2006) investigated a corpus of translated and original text from an Italian current affairs journal using a Support Vector Machine classifier, managing ca. 87% accuracy in distinguishing the two textual classes. Their study also investigated the performance of humans on such a task and found that the machine learning algorithm was more consistent although it was outperformed by one of the expert human analysts. Ilisei et al. (2010) used textual features such as type-token ratio and readability scores in their work on detecting translated text in Spanish and obtained comparable accuracy to Baroni and Bernardini (2006) who mostly used mixed POS and word n-grams. Popescu (2011) employed a different approach using a feature set consisting of n-grams of characters, and maintained reasonable accuracy in classifying translated literary works from originals.

Koppel and Ordan (2011) concerned themselves with the concept of *dialects* of translationese and whether translations from the same source language were more similar to one another than translations from different source languages and to what extent genre affected translationese. In their experiments on the Europarl corpus and a three source-language corpus from the International Herald Tribune, they found that training on one corpus and testing on another reported low accuracy, indicating genre effects, coupled with the fact that training on a corpus of translations from one source language and testing on a corpus translation from another source language obtained poorer results than using a corpus of translations from several source languages.

van Halteren (2008) investigated the predictability of source language from a corpus of Europarl translations and predicted source language with an accuracy of over 90%, using multiple translations of a source text in different languages. Distinguishing features from the Europarl corpus included phrases such as *a certain number* in texts of French origin, *framework conditions* in texts of a German origin and various features that were particular to the nature of the corpus as a collection of parliamentary speeches.[4] More recently, Lynch and Vogel (2012) revisited the source language detection task with a focus on literary translations, and obtained classification accuracy of ca. 80% on a corpus of translations into English from Russian, German and French using a feature set containing a combination of ratios of parts of speech and POS n-grams. Texts translated from French had a higher ratio of nouns to total words than the other two categories, and the frequency of contractions such as *it's* and *that's* varied between the subcorpora.

Focusing on the stylistic variation of individual translators from the point of view of researchers in translation studies, Baker (2000) defined frameworks for performing stylistic analyses of translator's using quantitative methods. Her own examples examined translators of Portuguese and Arabic and focused on the translation of common verbs, such as *say* and *tell*. She found that the frequency of these verbs was a distinguishing metric between translators but was careful to mention that these features might vary depending on the corpora in question. Winters (2007) profiled translator style in two translations of F. Scott Fitzgeralds *The Beautiful and the Damned*, focusing on modal particles and speech act reporting verbs as a distinguishing aspect of translatorial style. Vajn (2009) applied textual metrics such as type-token ratio and relative vocabulary richness to two translations of Plato's *Republic* to investigate the variation between two translations by Benjamin Jowett and Robin Waterfield and developed a theory of co-authorship to explain the complementary stylistic effect of authorial and translatorial style.

[3] See Li et al. (2011) and Wang and Li (2012) for examples of studies of translation from Chinese and English which take the cultural background of translators into account when discussing distinguishable features.

[4] German native speakers addressed the congregation in a different manner to English native speakers, for example.

Ongoing work in translation studies and digital humanities have examined the question of translatorial vs. authorial style using computational analyses. Burrows (2002) investigated the stylistic properties of several English translations of Roman poet Juvenal using his own Delta metric developed for authorship attribution and the frequencies of common words, Lucic and Blake (2011) investigated two translations of German author Rainer Maria Rilke in English using the Stanford Lexical Parser and found differing patterns of syntactic structure such as negation modifiers and adverbial modifiers.[5]

Recently, Forsyth and Lam (2013) analysed two parallel English translations of the French-language correspondence of Theo and Vincent Van Gogh using k-nearest neighbour classifiers and a feature set consisting of the sixty-nine most frequent words and found that a distinct authorial style for each of the brothers was preserved in both translations, with translatorial style also proving distinguishable, albeit to a lesser extent than its authorial counterpart. Lynch (2013) investigated two English translators of Henrik Ibsen's dramas using machine learning methods and found that document metrics and n-gram features similar to those used in this current study proved accurate in distinguishing authorship of parallel translations of the same source, and also that document metrics such as average sentence length distributions learned from translations of different works by the same author could be used to classify the author of a parallel translation, indicating that the translators' styles were learnable across a diverse corpus of works by the same author.

Rybicki (2006) used Burrow's Delta to investigate the stylistic nature of character idiolects in dramatic translation, focusing on Polish drama, and found that the translated idiolects tended to cluster in similar patterns[6] to the idiolects in the original text. Lynch and Vogel (2009) worked on a similar topic, the clustering of character idiolects in English and German translations of Henrik Ibsen's plays using the χ^2 metric. Rybicki and Heydel (2013) used Burrow's Delta, and dendrogram clustering to investigate the case of a Polish translation of Virginia Woolf's *Night and Day* and found that the method identified the point in the novel where one translator had taken over from another[7] Rybicki (2012) had previously used these techniques to distinguish translatorial style in a large corpus of Polish translations and concluded that such style was not to be captured using the methods at hand, which consisted of using Burrow's Delta metric with five thousand of the most frequent words. Although the metric performed well at clustering translations by author, it failed to cluster translations by translator, leading the author to conclude that as Venuti (1995) had claimed, the best translators are in fact invisible.

Although these studies are generally of an exploratory nature and often seek to draw conclusions about particular literary works and figures, the methodologies used are general to textual stylometry and have been successfully applied to emerging tasks in computational linguistics such as MT quality estimation, (Felice and Specia, 2012), personality detection (Mairesse and Walker, 2008), sentiment analysis (Gamon, 2004), fraud detection (Goel and Gangolly, 2012) and many other studies where textual analyses are pertinent.

3 Motivation and background to study

In this study, the translations of a literary translator of a number of different authors are examined in order to measure the extent to which authorial style is preserved by the translator in question. This analysis encompasses features represented by n-grams of words or POS tags and also stylometric metrics based on whole texts, such as type-token ratio, lexical richness and readability scores. Previous work (Rybicki and Heydel, 2013; Burrows, 2002; Rybicki, 2012; Koppel and Ordan, 2011) focused on lists of highly frequent words in their analysis of translations. By using supervised learning techniques, it is possible to investigate exactly which words are discriminating between author's idiolects in translation, regardless of frequency, together with abstract representations of word types and textual metrics, which present an alternative overview of the data in question.

This study examines the translations of British translator Constance Garnett (1861-1946) from the Russian originals written by Fyodor Dosteyevsky, Ivan Turgenev and Anton Chekhov. Moser (1988) and

[5] *not* and *nearly*.

[6] Villians with villians, heroes with heroes and female and male characters formed separate clusters

[7] The original translator passed away before she could finish the translation, hence the completion by another party.

Remnick (2005) write about Garnett's[8] life, describing her early days as a student of Latin and Greek in Cambridge, marriage to publisher and literary figure Edward Garnett and her chance introduction to Russian literature by the chance meeting with a young revolutionary in London. Along with the three aforementioned characters, she also translated works by Leo Tolstoy and Nikolai Gogol, Alexander Ostrovsky and Alexander Herzen, seventy works in all.

According to Moser (1988), her reputation was firmly established with her translations of Turgenev and thereafter Garnett was more or less responsible for igniting the English language-world's love affair with Dosteyevsky. Her translations were not without criticism however, Moser (1988) mentioning that Edmund Wilson believed she caused Russian authors to sound *more or less the same*, a claim echoed later by Joseph Brodsky who remarked that the average Western English-language reader cannot distinguish Tolstoy's voice from Dosteyevsky's, as they are in fact reading Constance Garnett's own voice.

Indeed, Remnick (2005) describes Garnett's translation style and mentions how she translated at breakneck speed, often skipping over sections which she did not understand. He also mentions Vladimir Nabokov's disdain for Garnett's translations, who was known to scribble vitriolic notes in the margins of Garnett translations during his tenure as a professor at Cornell and Wellesley in the United States. Remnick notes that children's book author Kornei Churnosky praised her translations of Turgenev and Chekhov but was less than pleased with her rendering of Dosteyevsky, complaining that she had smoothed over the erratic and challenging original text of that particular author. Thus, this work focuses on these claims of distinguishability in particular, for it is exactly these characteristics that can, in principle, be investigated using *supervised learning* techniques: Is it the case that one can automatically distinguish Garnett's renderings of Dosteyevsky from her translations of Turgenev, and if so, based on which textual characteristics, word distributions or individual word frequencies?

4 Corpus and methodology

The corpus was limited in these experiments to works by Dosteyevsky, Turgenev and Chekhov as these were the three authors translated by Garnett for which the most public domain text was available. Texts were downloaded from Project Gutenberg.[9]The final corpus consisted of eight works by Turgenev, seven works by Dosteyevsky and eleven collections of short stories by Chekhov. A selection of random text was made from each work matching the size of the smallest possible size of a work by each author and this selection was then divided into chunks of ten kilobytes each. The resulting corpus contains 942 segments from the three authors, 330 from Chekhov, 192 from Turgenev and 420 from Dosteyevsky. TagHelperTools was used to create the n-gram tokens, (Rosé et al., 2008) and calculate nineteen document statistics using TreeTagger, (Schmid, 1994) to tag texts for parts-of-speech. Weka, (Frank et al., 2005) was used for the *supervised learning* experiments, the SMO implemenation of a Support Vector Machine classifier along with the Naive Bayes and Simple Logistic Regression algorithms were used in the experiments.

The eighteen document level metrics used in the experiments are listed in Table 2. These were influenced by features used by Ilisei et al. (2010) in work which examined the problem of *translationese* detection in Spanish text. The two readability metrics employed are the Coleman-Liau Index, (Coleman and Liau, 1975) and the Automated Readability Index, (Smith and Senter, 1967). The n-gram features are calculated using TagHelper tools and the frequency of these features were reduced to a binary variable detailing the occurrence or non-occurrence of each feature in each segment.

5 Experiments

5.1 Document-level metrics

Experiments were carried out using different feature sets on the corpus described in Section 4. The experiments seek to classify the original author of a translated textual segment. The SVM classifier managed to achieve 87% accuracy when averaged using ten-fold cross validation on the whole corpus

[8](*nee* Black)

[9]www.gutenberg.org

Work	Author	Work	Author
The Bishop & O. Stories	Chekhov	The Cook's Wedding	Chekhov
The Chorus Girl	Chekhov	The Darling	Chekhov
The Duel	Chekhov	The Horse-Stealers	Chekhov
The School Master	Chekhov	The Party	Chekhov
The Wife	Chekhov	The Witch	Chekhov
Love & O. Stories	Chekhov	A Raw Youth	Dosteyevsky
Brothers Karamasov	Dosteyevsky	Crime & Punishment	Dosteyevsky
The Insulted and The Injured	Dosteyevsky	The Possessed	Dosteyevsky
White Nights	Dosteyevsky	Five Stories	Dosteyevsky
A House of Gentlefolk	Turgenev	Fathers & Children	Turgenev
On The Eve	Turgenev	Knock,Knock,Knock	Turgenev
Rudin	Turgenev	Smoke	Turgenev
The Torrents of Spring	Turgenev	The Jew	Turgenev

Table 1: Literary works in study

Feature	Desc.	Feature	Desc.
nounratio	nouns vs. total words	avgwordlength	average word length
pnounratio	pronouns vs. total words	prepratio	prepositions vs total words
lexrich	lemmas vs. total words	grammlex	closed vs. open class
complextotal	>1 verb: total sent.	simple complex	> 1 verb : <= 1 verb
simpletotal	<= 1 verb : total sent.	avgsent	average sentence length
infoload	open-class : total words	dmarkratio	discourse markers : total words
CLI	readability metric	fverbratio	finite verbs : total words
conjratio	conjunctions : total words	ARI	readability metric
numratio	numerals : total words	typetoken	word types : total words

Table 2: Document-level metrics used

using document-level features only. This result suggests that the authorial style of the three authors in question has indeed been preserved in translation.

Examining the features ranked by information gain in Table 3, it is clear that the ratio of nouns to total words and the ratio of pronouns to total words are highly distinguishing between the original authors. Ratio of prepositions to total words and the type-token ratio also feature in more elevated positions on the list than readability scores and sentence length measures.

5.2 N-gram features

The next set of experiments concerned the use of n-gram features, namely word unigram and POS bi-grams. For the word features, all noun features were removed as these, while providing clues to the identity of the author of a translation, are arguably not universal features of authorial style[10]. Verb features were not removed in such a fashion, however it may be argued that these also contain topical information and should be treated with caution. The remaining features were ranked by efficacy using the information gain metric and ten-fold cross validation and a subset of one hundred features were used for the classification experiments.

The SVM classifier in Weka with a linear kernel obtained 89.5% accuracy using a dataset of 100 words. The Simple Logistic regression classifier obtained 91.5% accuracy using the same feature set. This feature set was obtained by ranking the total list of word unigrams using information gain over ten-fold cross validation and removing the noun features as mentioned above. These high accuracy scores

[10]There is interest in lexical variation in translation, (Kenny, 2001) but this work focuses on stylistic features such as verbs and closed-class words as they are less prone to bias from the themes or topics in a text

obtained further reinforce the results obtained by using the document-level metrics, that a distinct textual style is learnable from the translations by Garnett of Dosteyevsky, Tolstoy and Turgenev. A number of these features and their relative frequencies are displayed in Table 6.

Feature	Rank.	Feature	Rank.
nounratio	1	avgwordlength	2
pnounratio	3	prepratio	4
typetoken	5	lexrich	6
simpletotal	7	simplecomplex	8
complextotal	9	grammlex	10
avgsent	11	infoload	12
cli	13	fverbratio	14
numratio	15	ari	16
conjratio	17	dmarkratio	18

Table 3: Metrics ranked using information gain and ten-fold cross validation

Feature set	Algorithm.	Accuracy
18 doc metrics	SVM	87%
18 doc metrics	Naive Bayes	74.2%
18 doc metrics	Naive Bayes	87.89%
100 words	SVM	89.5%
100 words	SimpLog	91.5%
1021 POS bigrams	SVM	83 %
1021 POS bigrams	SimpLog	78.98%
1021 POS bigrams	Naive Bayes	80%
1153 mix	SVM	95%
1153 mix	SVM	94.6 %
1153 mix	SimpLog	95%

Table 4: Accuracy overview

Using the 1021 unique POS bigrams which are present in the corpus as features, 83% classification accuracy was obtained using the SVM classifier, with Naive Bayes and Simple Logistic Regression managing 80% and 78.98% respectively.

5.3 Combined feature sets

Combining the feature sets from each of the experiments above, accuracy is improved. SVM obtains 95% accuracy, Naive Bayes and Simple Logistic Regression manage 94.6% and 95% respectively. This combined set contains 1153 features, 1021 POS bigrams, one hundred words and eighteen document level features. Ranking these features using ten-fold cross validation and Information Gain, the ranking displayed in Table 5 is obtained. Word unigrams and document-level features dominate the top fifty ranked features, with a number of POS-bigrams also occurring in the list.

6 Discussion

Tables 6 and 7 reflect the individual characteristics of each of the three authorial subcorpora examined here. The translations of Turgenev are distinguished by the higher average frequencies of the verbs *observed*, *repeated* and *replied*. Taking the value of the document-level metrics into account, Turgenev is to some extent unremarkable by these measures, although his works report higher average values for the two readability metrics, CLI and ARI, than the other two authors. The translations of Dosteyevsky distinguish themselves by the higher frequencies of adverbial forms such as *almost* and *perhaps*, which

Feature	Rank.	Feature	Rank.	Feature	Rank	Feature	Rank
prepratio	1	**pnounratio**	2	**nounratio**	3	almost	4
avgwordlength	5	observed	6	**simplecomplex**	7	**complextotal**	8
simpletotal	9	replied	10	repeated	11	near	12
smell	13	perhaps	14	**avgsent**	15	big	16
cried	17	added	18	sigh	19	rather	20
however	21	dark	22	purpose	23	sighed	24
certain	25	**typetoken**	26	**lexrich**	27	fact	28
few	29	eat	30	certainly	31	slowly	32
moment	33	**cli**	34	black	35	remarked	36
BOL_VBG	37	simply	38	ll	39	contrary	40
idea	41	quite	42	drank	43	CC_NNS	44
FW_NNP	45	NNP_RB	46	ah	47	high	48
ate	49	believe	50	slightly	51	**infoload**	52

Table 5: Mixed feature set ranked using information gain and ten-fold cross validation

Author	**almost**	**near**	**observed**	**perhaps**	**repeated**	**replied**	**smell**
Chekhov	0.000247	**0.000574**	0.000041	0.000289	0.000168	0.000013	**0.000226**
Dosteyevsky	**0.000958**	0.000244	0.000223	**0.001107**	0.000168	0.000042	0.000026
Turgenev	0.000741	0.000497	**0.000508**	0.000437	**0.000592**	**0.000373**	0.000070
Author	**added**	**big**	**cry**	**dark**	**however**	**rather**	**sigh**
Chekhov	0.000091	**0.000569**	**0.000361**	**0.000875**	0.000109	0.000146	**0.000757**
Dosteyevsky	0.000335	0.000131	0.000339	0.000285	0.000374	0.000446	0.000230
Turgenev	**0.000530**	0.000171	0.000198	0.000565	**0.000462**	**0.000538**	0.000483
Author	**certain**	**certainly**	**feat**	**fact**	**few**	**sighed**	**slowly**
Chekhov	0.000225	0.000114	0.003798	0.000482	0.000108	**0.000240**	0.000199
Dosteyevsky	**0.000763**	**0.000323**	0.003169	**0.000909**	0.000203	0.000022	0.000077
Turgenev	0.000576	0.000322	**0.004093**	0.000507	**0.000446**	0.000105	**0.000309**
Author	**I'll**	**black**	**contrary**	**idea**	**moment**	**remarked**	**simply**
Chekhov	0.014557	**0.000450**	0.000035	0.000312	0.000378	0.000001	0.000263
Dosteyevsky	0.013233	0.000170	**0.000205**	**0.000806**	**0.000999**	0.000014	**0.000701**
Turgenev	**0.016007**	0.000351	0.000093	0.000421	0.000355	**0.000140**	0.000342

Table 6: Relative frequencies for distinguishing words by author: Max values in bold

reflect uncertainty, but also adverbial forms such as *certain*, *certainly* and *simply*. They report a high average word length, and both a lower ratio of nouns to total words and lexical richness measure than the other two texts. They are not particular distinguished by their frequencies of verbal usage. The Chekhov translations are distinguishable by higher frequencies of *near* and *smell*, coupled with a lower average sentence length[11] and lower ratios of pronouns and prepositions to total words respectively. The three sentence type metrics are also distinctive. Perhaps the genre of the corpus has an effect here, as all of the included works by Chekhov are short stories while contributions from the other authors are primarily novels and novellas. Temporal variation or development of translatorial style may also play a role in any distinction, Garnett first began translating Turgenev in the late 19th century, followed by Dosteyevsky and Chekhov in the early 20th century, and it is probable that her knowledge of Russian and own writing style in English may have evolved over these years.

Reporting verbs[12] have been examined by Winters (2007), Mikhailov and Villikka (2001) and Baker (2000) in their work on finding distinguishing features of parallel translations of the same text. Here they

[11] Just over fifteen words, compared with over eighteen words for the other two authors.

[12] *Observed*, *repeated*, *replied* can be considered part of this category.

Author	**Chekhov**		**Turgenev**		**Dosteyevsky**	
Attribute	Mean	StdDev	Mean	StdDev	Mean	StdDev
grammlex	0.6553	0.0419	0.6388	0.0518	0.6849	0.0664
infoload	0.4482	0.0176	0.455	0.0237	0.4509	0.027
avgsent	**15.8881**	5.6812	18.6987	5.2877	18.1451	6.6252
nounratio	0.1759	0.0176	0.1723	0.0239	**0.1522**	0.0287
fverbratio	0.0903	0.0083	0.0942	0.0093	0.0943	0.01
pnounratio	**0.1047**	0.017	0.1184	0.0176	0.1278	0.0224
prepratio	**0.0423**	0.0071	0.0336	0.0057	0.0354	0.0068
conjratio	0.0913	0.0116	0.0867	0.011	0.0917	0.0148
numratio	0.0065	0.0025	0.0048	0.0021	0.0065	0.0036
typetoken	0.2954	0.0219	0.3007	0.0297	0.2758	0.031
avgwordlength	12.4996	0.6329	12.4417	0.7341	**13.4928**	1.0065
cli	3.8567	3.3722	5.4318	3.3074	5.0254	4.0703
ari	5.8797	0.9478	6.1201	1.1031	5.9542	1.2986
lexrich	0.2567	0.021	0.2586	0.0271	**0.2372**	0.0303
simplecomplex	**2.0079**	1.3585	1.278	0.526	1.3952	0.6163
dmarkratio	0.0011	0.0008	0.0015	0.0008	0.0012	0.0009
complextotal	**3.0075**	1.3584	2.278	0.526	2.3951	0.6162
simpletotal	**1.7428**	0.5121	1.9679	0.6037	1.9226	0.6496

Table 7: Mean and standard deviation per author: document metrics

occur as distinguishing features of authorial idiolects within works by the same translator. Of course, the efficacy of these features may be increased in these experiments as a result of eliminating noun features, although this was done in an attempt to mitigate the effect of topic based classification of the works of a particular author, and focus on features which represent deeper stylistic patterns. Further analyses of these phenomena must consult the nature of the source text, investigating to what degree of accuracy can the original works of each author be distinguished from one another.

7 Conclusions and Future Directions

This study has demonstrated the efficacy of supervised learning techniques as applied to the task of distinguishing authorial style in a literary corpus translated from Russian to English by a single translator. Both document metrics and n-gram features perform very well for this task, obtaining accuracies of over 80% using feature sets from each category. Combined feature sets improved performance, resulting in 95% classification accuracy between the three authors in question. Highly ranked features included the ratio of nouns to total words, the ratio of pronouns to total words and the ratios of prepositions to total words, also adverbs and reporting verbs such as *almost*, *observed*, *replied* and *repeated* and *near*. These results imply that in this case there is indeed a clear preservation of the individual authorial style by the translator in question, which to some extent refutes the claims of stylistic similarity or sameness across this particular translator's canon.[13], and supports the theory of a *translator's invisibility* as claimed by Venuti (1995). One aspect of the problem not focused on in this study is the relationship between the source and target text, and it is of interest in future work to investigate to what degree the stylistic shifts in translator's style reflect the original source text, or does the translator in fact create their own defined idiolect for a particular author? Further work may investigate how Garnett's style is distinct from another translator, there is evidence of stylistic differences existing between authors, and also between translators, with different features proving discriminating in both cases, as found in studies by Forsyth and Lam (2013) and Lynch (2013).

Future work on this topic will encompass a wider range of translators and languages in order to inves-

[13]Comments by Vladimir Nabokov and others as refered to by Remnick (2005).

tigate more general patterns in translated literature. Results using relatively shallow linguistic features such as POS n-grams and word class distributions have proven themselves useful in distinguishing authorial variation in a translator's style, however it is also of interest to apply deeper linguistic processing to these texts in order to investigate more fine-grained elements of authorial and translatorial style within text. Examples of technologies which could be applied include semantic role labeling, (Swier and Stevenson, 2004) deep syntactic parsing, (Lucic and Blake, 2011), and LDA for detecting levels of metaphor (Heintz et al., 2013), in order to obtain a clearer picture of the stylistic structure of such documents.

Acknowledgements

The Centre for Applied Data Analytics Research is an Enterprise Ireland and IDA initiative. Many thanks to Dr. Daniel Isemann at Universität Leipzig for comments on an early draft of this work and to Prof. Carl Vogel at Trinity College Dublin who provided guidance, inspiration and extensive comments on previous studies in this space.

References

M. Baker. 2000. Towards a methodology for investigating the style of a literary translator. *Target*, 12(2):241–266.

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

L. Borin and K. Pruetz. 2001. Through a glass darkly: Part-of-speech distribution in original and translated text. *Language and Computers*, 37(1):30–44.

J. Burrows. 2002. The Englishing of Juvenal: computational stylistics and translated texts. *Style*, 36(4):677–699.

Meri Coleman and TL Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Mariano Felice and Lucia Specia. 2012. Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103. Association for Computational Linguistics.

Richard S. Forsyth and Phoenix W. Y. Lam. 2013. Found in translation: To what extent is authorial discriminability preserved by translators? *Literary and Linguistic Computing*.

E. Frank, M.A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, and I.H. Witten. 2005. Weka: A machine learning workbench for data mining. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, pages 1305–1314.

Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics.

Sunita Goel and Jagdish Gangolly. 2012. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management*, 19(2):75–89.

Ilana Heintz, Ryan Gabbard, Mahesh Srinivasan, David Barner, Donald S Black, Marjorie Freedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphor with lda topic modeling. *Meta4NLP 2013*, page 58.

I. Ilisei and D. Inkpen. 2011. Translationese traits in romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*.

I. Ilisei, D. Inkpen, G. Corpas Pastor, and R. Mitkov. 2010. Identification of Translationese: A Machine Learning Approach. *Computational Linguistics and Intelligent Text Processing*, pages 503–511.

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

M. Koppel and N. Ordan. 2011. Translationese and its dialects. *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA*.

D. Kurokawa, C. Goutte, and P. Isabelle. 2009. Automatic Detection of Translated Text and its Impact on Machine Translation. In *Proceedings of the XII MT Summit,Ottawa, Ontario, Canada*. AMTA.

G. Lembersky, N. Ordan, and S. Wintner. 2011. Language Models for Machine Translation: Original vs. Translated Texts. *Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 2011*.

D. Li, C. Zhang, and K. Liu. 2011. Translation style and ideology: a corpus-assisted analysis of two english translations of hongloumeng. *Literary and Linguistic Computing*, 26(2):153.

Ana Lucic and Catherine Blake. 2011. Comparing the similarities and differences between two translations. In *Digital Humanities 2011*, page 174. ALLC.

Gerard Lynch and Carl Vogel. 2009. Chasing the ghosts of ibsen: A computational stylistic analysis of drama in translation. In *Digital Humanities 2009: University of Maryland, College Park, MD, USA*, page 192. ALLC/ACH.

Gerard Lynch and Carl Vogel. 2012. Towards the automatic detection of the source language of a literary translation. In Martin Kay and Christian Boitet, editors, *COLING (Posters)*, pages 775–784. Indian Institute of Technology Bombay.

Gerard Lynch. 2013. *Identifying Translation Effects in English Natural Language Text*. Ph.D. thesis, Trinity College Dublin.

F. Mairesse and M. Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 165–173.

M. Mikhailov and M. Villikka. 2001. Is there such a thing as a translators style? In *Proceedings of Corpus Linguistics 2001, Lancaster, UK*, pages 378–385.

Charles A Moser. 1988. Translation: The achievement of constance garnett. *The American Scholar*, pages 431–438.

M. Popescu. 2011. Studying translationese at the character level. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP'2011). Hissar, Bulgaria*.

David Remnick. 2005. The translation wars. *The New Yorker*, 7:98–109.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271.

Jan Rybicki and Magda Heydel. 2013. The stylistics and stylometry of collaborative translation: Woolfs night and day in polish. *Literary and Linguistic Computing*, 28(4):708–717.

J. Rybicki. 2006. Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. *Literary and Linguistic Computing*, 21(1):91–103.

J. Rybicki. 2012. The great mystery of the (almost) invisible translator. *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*, page 231.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

EA Smith and RJ Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, page 1.

Robert S Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of EMNLP*, volume 95, page 102.

Dominik Vajn. 2009. *Two-dimensional theory of style in translations: an investigation into the style of literary translations*. Ph.D. thesis, University of Birmingham.

H. van Halteren. 2008. Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944. Coling 2008 Organizing Committee.

L. Venuti. 1995. *The translator's invisibility: A history of translation*. Routledge.

Q. Wang and D. Li. 2012. Looking for translator's fingerprints: a corpus-based study on Chinese translations of Ulysses. *Literary and Linguistic Computing*.

Marion Winters. 2007. F. scott fitzgerald's die schönen und verdammten: A corpus-based study of speech-act report verbs as a feature of translators' style. *Meta: Journal des traducteurs*, 52(3).