Strange bedfellows: Shifting paradigms in the corpus-based analyses of literary translations

Gerard Lynch
Centre for Applied Data Analytics Research
Department of Computer Science and Informatics
University College Dublin, Ireland
firstname.lastname@ucd.ie

Abstract

Although the application of computational and statistical approaches in the analyses of literary translations has gained considerable traction of late, there is still a good deal of progress to be made on bridging the divide between the cross-disciplinary nature of such analyses and the traditions of translation studies and the computing sciences. This article attempts to document issues and methodological divergences which arise in such work, collaborative or otherwise, drawing upon studies across the digital humanities, computational stylometry, corpus linguistics and translation studies literature and identifying areas in which further progress could be achieved based on a mutual collaborative spirit between disciplines.

1 Introduction

The application of computational and/or statistical analyses to the written word is a practice that dates back to as early as the 19th century, with Mendenhall (1887) pioneering word-length distribution analysis of British authors Dickens and Thackeray published in Science, followed by an expanded analysis which compared Shakespeare with his contemporaries in Mendenhall (1901). This experiment was inspired by personal correspondence from British logician Augustus de Morgan a number of years prior who touched on the idea of using word length measures to determine the author of the Pauline epistles. From these inauspicious steps towards solving a number of age-old questions grew a field which now spans disciplinary boundaries. Zipf (1935) defined the eponymous law motivating the field of corpus linguistics, stating that the frequency of any word in a natural language corpus is inversely proportional to its rank, although this natural phenomenon had also been noted prior in relation to population growth in cities by Auerbach (1913).

British statistician George Udny Yule examined sentence length distributions in the works of Coleridge and Bacon in Yule (1944) and carried out an attribution exercise on the works of Thomas a Kempis, spearheading a second wave of authorship attribution exercises using statistical methods. After the Second World War, a number of new contributors ushered in a new age of computational language analysis, the value of Father Roberto Busa’s work, the first digital humanist and pioneering collaboration with Thomas Watson from IBM on a digital concordance of texts by Thomas Aquinas and the work of Warren Weaver on the possibility of a system facilitating automated translation of Russian into English, a Cold War-era US intelligence project that would create the field of machine translation which would generate much of the research methodology and fund the resources firmly establishing computational linguistics and natural language processing as a discipline in the twentieth century. Complementing this pioneering work came a study which swung the attention back to the age-old task of authorship attribution, namely the work of Mosteller and Wallace (1963) on the
authorship attribution of the Federalist papers, the first such work to use a computer to analyse texts, previous works relying on statistical methods which were computed by hand.

In today’s era of Google Translate, *culturnomics* and *big data*, one could imagine that computational analyses of literary translation would have become straightforward and widespread, but this is not necessarily the case as evidenced by the literature. Although the application of corpus linguistic methods and concordance have become commonplace in translation studies since the 1990’s, there appears to be a smaller body of work concerned with applying the state-of-the-art in natural language processing and linguistic analyses to answering questions about literary translations in particular. There are a number of factors which may be hampering research collaborations such as these and these factors are outlined in the following article, along with a number of suggestions for possible future collaborative studies.

2 Trends across disciplines in translation stylometry

This section attempts to contrast the methodological treatment of corpus studies in literary stylometry from the perspective of translation studies and corpus linguistics, the digital humanities and literary stylometry and the new emergent paradigm which employs text classification and machine learning approaches for the purposes of carrying out research in the area.

2.1 First phase: Of translation universals and early corpus studies

Since Gellerstam (1986) coined the term *translationese* as a descriptor for what Frawley (1984) referred to as the *third code* and to some extent a related phenomenon to the *interlanguage* defined by Selinker (1972) in the language acquisition context, there has been interest in the application of corpus linguistic methods to questions of translation stylometry, in particular quantifying any distinctions between translationese and non-translated language. (Baker et al. (1993), Baker (2000)) defined a framework for investigating the stylistic properties of parallel translations, giving examples from the literary domain. Although not employing detailed computational methodology, the analysis of word frequencies was not common at the time in the translation studies community. She posed three main questions when investigating stylistic variation in literary translation, whether a translator’s preference for particular linguistic constructions was independent of the authorial style of the original author, whether the stylistic choices are independent of the general preferences of the target language and if so, could they be explained by knowledge about the social, cultural or ideological position of the translator.

Adopting similar methodology, Mikhailov and Villikka (2001) focused on Finnish translations from Russian using statistical measures from corpus linguistics, reporting consistencies in translator’s style across separate translations. Their approach results in the conclusion that modals and sentence length define translator’s style, and they do not consider the translator’s social background or other reasons for the stylistic variation. Olohan (2001) focuses on the explicitation translation universal in her work which focused on optional items in translated English, comparing a corpus of translations with the British National Corpus using statistical tests. Following these studies were substantial corpus studies on lexical choice in translation, (Kenny 2001), simplification, (Laviosa-Braithwaite 1997) and many others, which have been excellently summarised previously by Zanettin (2013).

More recently, Saldanha (2011) expands upon Baker’s blueprint for corpus linguistic investigations in translation studies, investigating non-parallel translations from Portuguese and Spanish by Peter Bush and Margaret Jull Costa with a focus on stylistic features such as italicized text, preservation of cultural expressions from the source language and user of the that connective. The work also involves comparing the individual frequencies of each feature per translator with a bilingual reference corpus in order to situate the feature variation in a larger linguistic context. Bosseaux (2006) investigated the translators’ divergent choice of
second-person personal pronouns in the two French translations of *The Waves* by Virginia Woolf. Winters (2007) work on German translations of F. Scott Fitzgerald’s The Beautiful and the Damned identified speech-act reporting verbs as markers of translator’s style, and that one translator repeats speech act reporting verbs more faithfully with the source. Li, Zhang, and Liu (2011) investigate translator style from a more sociological perspective, contrasting the translations of the Chinese epic Hongloumeng by a British sinologist and an official Chinese government-sanctioned translator, focusing on type-token ratio and related corpus measures to profile stylistic divergence and thereby translation ideology. This topic is continued by Huang and Chu (2014) in their comparison of the translations of a prolific Chinese-English translator, Harold Goldblatt with those of Gladys Yang.

As evidenced by the literature, these studies tend to be fine-grained accounts of the translation of one particular work or of a pair of translators, paying close attention to any cultural or ideological bias or effects in the translation which can be attributed to the translator. This can contrast somewhat with the studies in literary stylometry in the next section, which can be indeed fine-grained but tend not to focus on cultural and ideological markers in translation, rather issues of stylistic development over time, influence and identifying divergent patterns in joint translations.

2.2 Second Phase: Burrowing into Translation and Corpora++

Alongside the research in the translation studies literature, scholars in the areas of digital humanities and literary stylometry have been applying corpus linguistic methods to questions of stylometry for quite some time. Although translations were not the main focus of the discipline, there are a number of scholars who focused on tasks in this space. Commonalities in these studies include the usage of non-standard statistical metrics such as John Burrow’s Delta method and a focus on the stylometric distinction of character idiolects in an author’s canon, which is often extended to translation. Another aspect of studies in this field which refer to translation is the focus on pre-20th century texts, presumably for reasons of data accessibility. Following on from his pioneering study on the analysis of character idiolects in the works of Jane Austen, (Burrows 1987), later work sees Burrows (2002b) inadvertently establishing a breakaway subfield of literary translation stylometry by applying his Delta metric to twenty translations from Latin to English of Roman poet Juvenal, identifying a seventeenth century translation by Thomas D’Urfey as the most similar to all of the other translations. Burrows concludes that this translation may have been used as a reference translation for later works. The application of stylometry to non-prose text is generally less common, although Pantopoulos (2009) investigates the stylistic characteristics of four English-language translators of modern Greek poetry using corpus linguistic methods and recent work by Herbelot (2014) investigates a distributional semantics approach to modelling poetry.

Rybicki (2006) takes up the mantle of literary translation stylometry and applies Burrow’s Delta metric to character idiolects in translations of the Polish epics of Sienciewicz and finds that similarly distinct clustering of character idiolects are observed in the translation as well as the original text. Lynch and Vogel (2009) apply Rybicki’s methodology coupled with the $\chi^2$ metric to translations of Henrik Ibsen in English and German and the original texts and observes similar patterns of distinctive characterisation, both in the original texts and in the translations. Rybicki (2012) continues his work on translation stylometry with Burrow’s Delta, finding that texts tended to cluster by author rather than translator in a corpus of English to Polish literary translations using frequent words as features, a result which he claims verifies Venuti (1995) and his theory of a translator’s invisibility. Later work by Rybicki and Heydel (2013) investigates a collaborative translation between two translators working on the same translation of Woolf’s Night and Day into Polish, where one translator completed the work of the other after her death, successfully identifying the point where the new translator took up the task. Related work in the Slavic language domain is carried out by Grabowski (2011,
2013) who investigates translation patterning in English, Russian and Polish versions of Nabokov’s Lolita, focusing on sentence length and word type/token distributions in source and target texts.

The majority of stylometric studies on literary translation use relatively shallow textual features however a handful of studies go beyond word n-grams and Burrow’s Delta in their scope. Lucic and Blake (2011) use parses from the Stanford Lexical Parser\(^8\) of two English translations of Rainer Maria Rilke and observe differing patterns of negation and adverbial modifiers between the two translations. A novel approach is employed by El-Fiqi, Petraki, and Abbass (2011) who use a network-theoretic linked representation of a text to identify patterns in two translations of the Holy Q’uran into English, Hung et. al. (2010) who use variable-length grams to attribute Chinese translations of 4th century Buddhist texts written in Sanskrit and Popescu (2011) investigates translations in a corpus of literary text from Project Gutenberg using string kernels\(^9\) as distinguishing features but advocates caution in this approach from an interpretability point of view. Recent work by Hedegaard & Simonsen (2011) investigated authorship using frame semantic representations, finding that these representations improved the attributions of translated texts, however did not perform as well on texts which had not been translated. Claesen (2014) focuses on coordination patterns in two translations of Joseph Roth’s Hotel Savoy from German to Dutch.

The literary stylometric studies which focus on translation have some elements in common with the studies from the translation studies literature, such as the focus on individual translators and authors, although the focal points of the research projects can often differ, with more of an emphasis on attribution, stylistic profiling and language evolution. The next phase of studies share properties including macro-level analysis and discovery of over-arching trends and universals when applied to translated text.

### 2.3 Third Phase: The Machine (Learning) Age

The adoption of machine-learning approaches to all possible hypotheses in the current era of big data does not exclude questions of corpora and translation and this phase follows the pioneering work of Baroni and Bernardini (2006) on a comparable corpus of Italian journalistic text. These methods can identify stylistic patterns in textual corpora with considerable ease and may be useful to identify hitherto unconsidered textual variation. These studies often employ a wide range of textual feature representations, including part-of-speech tokens, corpus statistics such as type-token ratio and lexical richness measures and word n-grams. Commonalities in studies employing such methodology include investigation of the stylistic patterns of individual translators and macro-analyses, often eschewing literary genres for journalistic text and focusing on translations as target language corpora only.

Questions regarding gender effects in translation stylometry have been raised in Shlesinger, Koppel, Ordan, and Malkiel (2010) following on earlier work by Koppel, Argamon, and Shimoni (2002) which investigates gender differences in literary language using machine learning methodology and Leonardi (2007) who investigated differences in translator ideology and language by gender. In the experiments on translated text, their classifiers failed to distinguish between male and female translators using frequent word features, with classification results barely above the chance baseline of 50%, although earlier corpus statistics had identified statistically significant features in the overall sub-corpora. The corpus used in these experiments consisted of 213 literary extracts translated from more than 30 source languages\(^10\). From this study, the authors surmise that the application of supervised learning approaches to questions of textual stylometry provide a more rigorous methodology for establishing the efficacy of stylistic features. Conventional statistics identified a set of features which occurred statistically more often in translations from translators of each gender however these features ultimately did not prove effective at discriminating the gender of the translator of a text using machine learning methods. The subject of language, gender and translation is one which deserves further investigation, perhaps using a source-language
restricted corpus to restrict variability and improve the efficacy of any machine learning experiments by restricting the number of possible factors of influence on textual style.

Lynch and Vogel (2012) apply support vector machine classifiers to a corpus of literary translations in order to detect the source language of a textual segment, an approach which had been considered previously for non-literary text such as Europarl and journalistic articles from the New York Times and Haaretz, among others (van Halteren, 2008, Koppel and Ordan, 2011). This approach succeeded in distinguishing the source language of a text and identified a number of corpus features and word and POS n-grams which were distinctive between translations from different source languages, however the study did not rigorously examine whether features were truly source language markers or somehow artefacts of the texts chosen. Recently, this work was replicated (Klaussner, Lynch, and Vogel, 2014) using a completely disjoint set of contemporaneous texts and comparable results were reported, along with additional distinctive features such as trigrams of part-of-speech tags which had not been employed in the previous study, which allays any fears about the original results being dependent on the particular corpus used.

More recent research using advanced machine learning methods challenges Venuti’s theory of a translator’s invisibility, Forsyth and Lam (2013) investigate parallel translations of the Van Gogh letters and identify separate stylistic traces of original authorship and translator’s style using both machine learning and corpus linguistic approaches in two parallel translations of the letters into English, finding that translator distinctiveness was less strong than authorial discriminability. This correlates with Vajn (2009) who developed a theory of two-dimensional style as applied to parallel English translations of Plato’s Republic, incorporating authorial and translatorial stylistic features into the descriptive process. Bogdanova and Lazaridou (2014) develop a related study investigating the preservation of authorial style across translation from an authorship attribution perspective and attempt to cluster works by their original author both in the original language and translation, one approach being to translate the translations back to the original language using machine translation and applying stylistic clustering techniques to the machine translation and the original text. An investigation of translator-style on parallel translations in Lynch (2013) identifies metrics such as average sentence length and type-token ratio as discriminatory on translations of Ibsen’s drama, along with frequent words such as because, nearer and recollect. Lynch (2014) focuses on the translations of Victorian-era British translator Constance Garnett and uses support vector machine classifiers to guess the original author of a translation segment by Garnett, obtaining high classification accuracy between the source authors using document statistics on the textual segments and even better accuracy employing linguistic features such as reporting verbs and part-of-speech clusters. This work was motivated by a article by Remnick (2005) on the translation of Russian literature in which Vladimir Nabokov and other contemporaries were quoted as mentioning that the translations of Garnett had no distinct voice, her translations of Turgenev being indistinguishable from her renderings of Dostoyevsky. On the subject of Nabokov, very recent work by Nisioi (2015) attempts to distinguish between Nabokov’s translated and original works, using an unsupervised learning approach and a list of closed class words as features, remarking that it was often difficult to detect stylistic differences between author and “author-as-translator”, as Nabokov translated many of the works himself from Russian to English.

3 Methodological and motivational considerations

3.1 Supervised learning and corpus tools

The state-of-the-art in text analytics research and analysis tends towards supervised learning approaches such as support vector machines, (Joachims 1998), Naive Bayes classifiers and neural networks. These have an extra degree of complexity than previous statistical metrics
employed in corpus linguistics however they can be very useful for discovering unknown patterns in data, when compared with traditional statistical measures such as t-tests and chi-square tests which require prior domain knowledge about which features to measure. Although there are both open-source and proprietary software packages available to carry out these analyses such as R\textsuperscript{17}, WEKA\textsuperscript{18}, and RapidMiner\textsuperscript{19}, these tend to come with a learning curve which may be too steep for curious humanists and corpus linguists who have firmly adopted software packages such as the IMS Open Corpus Workbench detailed in Christ (1994) and the WordSmith tools package from Scott (1996) for use in their descriptive analyses.

There is a growing need for more straightforward machine learning platforms which can be used by a greater subsection of the population at large, and Google’s Prediction API\textsuperscript{20} is a step in this direction, facilitating the application of machine learning for non-experts, although it does not focus specifically on text processing. The emergence of graphical tools which allow the creation of text-processing pipelines through a drag-and-drop interface is also a development of note in this space which can aid research in advanced textual stylometry in general. Perhaps a hybrid system combining a collocation-viewer with a graphically controlled machine learning module with language analytics functionality via the Stanford NLP Toolkit (Manning et. al. 2014) would be the ultimate general-purpose tool of choice for future studies in translation stylometry, although open-source solutions such as AntConc (Anthony et al. 2013) are approaching this level of functionality.

3.2 Corpora and data management

The field of statistical machine translation applies statistical methods to large parallel translation corpora such as Europarl and the Canadian Hansard corpus in order to derive models to translate from one language to another. Machine translation has been a mainstay of natural language processing research since Weaver’s pioneering paper in the 1940’s and attracts a considerable amount of research funding year-on-year. As a result, a good deal of computational stylistic analysis of translations focuses on corpora such as these, with scant regard for literary translations. Thus, large scale parallel corpora of literary translations are not as prevalent, with the translation studies literature focusing on small-scale analyses of one work or at most, the complete works of a particular author. As a result, studies which focus on a fine-grained analysis of the work of one author using computational methods can often be subject to critique in the computational linguistics literature, as the tendency is to use as large a corpus as possible to obtain statistically significant experimental results. Another aspect hampering fully parallel analyses of literary translations is the availability of parallel reference translations in online sources such as Project Gutenberg\textsuperscript{21} or WikiSource\textsuperscript{22} which tries to improve matters by imposing a wiki-format browsing structure on copyright-expired source texts. Given the effort that can be involved in digitising a print work, it is perhaps unsurprising that, copyright restrictions notwithstanding, it is unusual to find two contemporary translations of the same work on a site such as Project Gutenberg, unless that work is indeed of high literary regard. Due to this accessibility bottleneck, curious computational humanists can quickly become discouraged when searching for corpora for their investigations.

Cheesman, Thiel, Flanagan, Geng, Ehrmann, Laramee, Hope, and Berry (2010) propose the notion of translation arrays, encompassing linked databases of multiple translations of the same source text in various languages, arguing that such data would provide a rich resource for cross-cultural studies in language evolution, in the monolingual case\textsuperscript{23}, and general cultural questions also.
Their own related investigations have focused on the visualisation of textual re-use in translations of common texts such as the Bible, the same text examined by Covington, Potter, and Snodgrass (2014) who cluster Biblical translations using dendogram clustering and document-level metrics such as sentence length, average type-token ratio and “idea density”.

This approach can be extended to any texts for which multiple parallel translations are available.

3.3 Asking questions

Based on the literature surveyed here, the scholar’s own background can tend to influence how they approach a study in translation stylometry, based on the individual norms of their host discipline.

Applying best practices for computational studies can bias the approach of a research project in literary translation stylometry, such studies tend to adopt a macro-level view into translation stylometry, such as defining characteristics of translationese in a large corpus of reportage (Baroni and Bernardini 2006, Koppel and Ordan 2011), quantifying the existence of translation universals in technical and medical translations (Ilisei and Inkpen 2011) and separating a corpus of technical translations by translation direction (Kurokawa, Goutte, and Isabelle 2009).

Research stemming from the literary studies and translation studies domain tends to take a more fine-grained approach, usually attempting to ground theories about the cultural background or processes of the translator through an examination of their translation result, often compared with the source, following work by such scholars as Baker and Saldanha. Thus, the application of machine learning and advanced stylometric analysis can often be hampered by the relatively small size of the corpus under examination or the particularly fine-grained nature of the research question. On the other hand, those in the translation studies field often have first-hand unfettered access to contemporaneous copyrighted translations which are not available to the general public such as the Translational English Corpus, Olohan (2002), which is a considerable advantage in carrying out corpus studies. Thus, cross-discipline collaborations often produce the most interesting experimental design and studies, examples include projects by computational linguists Marco Baroni and translation studies scholar Silvia Bernardini and collaborations between translator and English literature professor and translator Jan Rybicki with corpus linguistics scholar Maciej Eder which can draw on strengths from individual fields and also respect the validity of investigatory norms of the disciplines.

4 Towards a shared future

With this spirit in mind, there are many areas in which both researchers of a quantitative nature and translation studies scholars may collaborate which maximise the skills and talents of each discipline. Within the computational linguistics community, there is a growing interest in the analysis of literary text, including translation, with the establishment of the Workshop on Computational Linguistics for Literature, currently in its fourth iteration, and issues of collaboration with the humanities have been documented by Hammond, Brooke, and Hirst (2013). With the relentless drive towards data-driven methods in the digital humanities, it is important not to lose sight of the original tenets of translation studies mainstays such as Baker who espouse that corpus-based studies should seek to ground their analyses in the cultural background, bias or preferences of a translator or movement. At the same time, progress in automatic text analytics has resulted in a set of tools and processes which enable detailed, semi-automated linguistic and stylistic analyses of large textual corpora in near real-time.

Neuman et.al. (2013) describe computational methods toward the analysis of the quantity of metaphor in text, such approaches may also be applied to literary translations in order to quantify the level of metaphor within these and also apply quantitative approaches in concert with theories of metaphor in translation such as the work by Steiner (2002). Also, any
application of Latent Dirichlet Allocation or topic modeling approaches, (Blei, Ng, and Jordan 2003) to translated text may serve to illustrate over-arching trends of interest within parallel translations and may prove a fruitful area of interest for the future.

Other topics which have not received substantial attention to date are questions related to how a translator’s style relates to their own authorial style. There is of course a rather limited pool of translators who are also published authors in their own right, however Wang and Li (2012) focus on this issue in their study of two Chinese translations of Joyce’s Ulysses, separating translation effects into those from the source language, such as the post-positioning of adverbial phrases in the target text from individual lexical choice by an individual translator, manifested in the choice of a dialectically sensitive translation of the verb to know in English and one translators systematic overuse of the Chinese verb duo25 in both his translation and his original writing.

Furthermore, the application of syntactic and semantic parsing to literary translations is another area which has received little attention, save previously mentioned work by Lucic and Blake (2011). One possible reason for this gap may be misconceptions about the applicability of software trained on modern textual corpora such as the Wall Street Journal to literary texts26, although some basic investigations can confirm resources are available for textual parsing in Latin, (Bamman and Crane 2006), medieval Portuguese, (Rocio, Alves, Lopes, Xavier, and Vicente, 2003) and medieval French (Mazziotta 2010), (Stein and Preˇvost 2013). In a similar vein, van Dalen-Oskam (2012) uses open-source general purpose named-entity recognition tools in her study of the translation of proper names in Dutch novels and their English translations, finding the accuracy of the NER software to be more than sufficient for the needs of the study, although mentions that some customisation of the software in order to recognise subtypes of proper names would be a welcome addition.

Many collaborative endeavours between humanists and computational scientists are commonly built around the idea of creating shared resources as evidenced by the Text Encoding Initiative and related projects27 although there is a rise in the number of data-driven humanities projects and a number of key scholars in the area span both fields quite comfortably, including Matt Jockers who has investigated large-scale topic models on thematic trends in literature, (Jockers 2013) and the team behind the Google Ngrams project, (Michel et. al. 2011). A possible study in this space could apply big data analysis methodology to a large corpus of parallel translation with a constant L1 and L2 to investigate trends in stylistic variation that may transcend individual translator’s choices.

Notes

1 Mendenhall mentions De Morgans 1872 work A Budget of Paradoxes in the introduction to his 1887 paper, although is himself unsure of the reference location.
2 A more complete history of computational text analysis in this period is given by Hockey (2004).
3 Michel et al. (2011) being the canonical work discussing the Google-Ngrams corpus and some analyses.
4 Candel-Mora and Vargas-Sierra (2013) and Zanettin (2013) discuss corpus linguistic studies in translation studies literature.
5 a dialect or variant of a language consisting solely of translations from other languages.
6 See Burrows (2002a) for a detailed explanation of the methodology with examples.
7 villians vs heroes, male vs. female reference.
8 Sequences of characters
10 These were obtained from the website translationswithoutborders.org.
11 Examples included readability scores and contractions such as that’s and it’s.
12 The correspondence was between Vincent Van Gogh and his brother Theo.
13 The stylistic divergence between the parallel translations.
14 The stylistic divergence between Theo and Vincent.
15 Russian authors Ivan Turgenev, Anton Chekhov and Fyodor Dosteyevsky.
16 Average sentence length, lexical richness etc.
17 http://www.r-project.org/.
18 Frank, Hall, Holmes, Kirkby, Pfahringer, and Witten (2005).
References


