Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 52 (2018) 79–104

www.elsevier.com/locate/csl

The translator’s visibility: Detecting translatorial fingerprints in contemporaneous parallel translations☆

Gerard Lynch*,[a], Carl Vogel[b]

[a] *School of Informatics, University College Dublin, Dublin, Ireland*
[b] *Computational Linguistics Group, School of Computer Science and Statistics, Trinity College Dublin, Ireland*

Received 29 July 2015; received in revised form 4 May 2018; accepted 6 May 2018
Available online xxx

Abstract

We detail the results of experiments towards a fine-grained stylometric analysis, the identification of distinguishing features between contemporaneous literary translations, both parallel works and also translations of non-parallel sets of works by the same author. We examine translations of plays by the Norwegian dramatist Henrik Ibsen with the initial point of focus being the Ibsen drama *Ghosts*, for which there exists comparable contemporaneous translations by R. Farqhuarson Sharp and William Archer. Consequently, a number of prose translations of Russian author Anton Chekhov by Marian Fell and Constance Garnett are examined in order to validate hypotheses formed from the results of the Ibsen study and investigate possible particularities in translator’s style which may vary according to genre.

By carrying out an analysis of these texts using a variety of machine learning approaches such as Support Vector Machines, Simple Logistic Regression, Naïve Bayes and Decision Tree classifiers, a number of distinguishing textual features are obtained, and the relative frequency of these features in the texts are compared to their frequencies in reference corpora in order to establish which features can be attributed to stylistic choices by the translators themselves and which features may be due to influence from the source language or the topic or genre of a text. We also use the popular Delta metric from authorship attribution studies to investigate the clustering of texts based on most frequent words and a list of discriminatory terms learned in the supervised machine learning experiments.

We find that common word unigrams and bigrams are the most salient features for translator fingerprinting across our two authors and four translators examined and are ultimately successful in our goal of classifying which text originated from a particular translator with accuracy measurements of over 90% on average.

Keywords: Stylometry; Authorship; Translation

1. Introduction

1.1. On textual stylometry

This paper details a study in textual stylometry, a subfield of computational linguistics which focuses on fine-grained stylistic classification of text. Related tasks include *authorship attribution, demographic profiling* and

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.

E-mail address: gerard.lynch@ucd.ie (G. Lynch), vogel@cs.tcd.ie (C. Vogel).

https://doi.org/10.1016/j.csl.2018.05.002

detecting *bias* in text. It is distinct from the more mainstream research area of *text classification* in that the features examined in *stylometry* are more often not content words such as *nouns* and *verbs* which convey meaning and topical signals, but rather stylistic elements such as pronouns, prepositions, conjunctions and other closed-class words often discarded during the pre-processing stage of general text classification. It is through the scrutiny of patterns of these tokens in which we may find evidence of stylistic properties, often unconsciously conveyed by the author.

Some examples of these stylistic patterns might be the tendency for a translator to prefer the perfect rather than imperfect tense in their translations, for example into English, which could be identified using frequencies of *part-of-speech tags*. Another consistent stylistic pattern may be related to the concept of *translation universals*, for example translator A tends to constantly expand upon the source text material in his or her translations, a tendency manifested in a longer average sentence length. One of the main challenges in any research into these questions are the confounding factors in textual stylometry, most commonly textual *period, source language* and also *genre*.

1.2. Research questions

The study itself focuses on questions pertaining to the *visibility*[1] of *translator style*, that is to say, given two parallel translations of the same text into the same language by two translators, what stylistic characteristics can we identify which differ between the translations and can these characteristics be attributed to translator's individual language preferences or some other phenomena?

These stylistic variations will be quantified both in the *frequency differences* between certain common words, but also in the variation in values of commonly used statistical measures of textual style such as *readability scores* and *average sentence length*. Furthermore, we attempt to quantify the *consistency* of a translator's style, providing the author remains constant, will a translator's translation of work *A* contain stylistic similarities to their translation of work *B*?

The concept of *translator style* has not attracted as much attention from researchers in the computational stylistics space based on our extensive review of the literature. Researchers tended to focus on more coarse grained classification tasks such as *source language detection* and comparing translated and non-translated text in the same language although some attempts have been made by literary scholars to identify features of translator style. The lack of interest in the topic is possibly due to the myriad of influencing factors possible when investigating stylistic characteristics of individual translators. Nevertheless, we consider it a fertile area for the application of methods which have already proven successful towards providing quantitative answers to related qualitative questions from translation stylometry.

2. Related work

2.1. Computational linguistics

Several studies in computational linguistics (Baroni and Bernardini, 2006; Pastor et al., 2008; van Halteren, 2008; Kurokawa et al., 2009; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Koppel and Ordan, 2011; Popescu, 2011; Lynch and Vogel, 2012; Volansky et al., 2013; Klaussner et al., 2014) have focused on the computational analysis of translated text. Baroni and Bernardini (2006) were concerned with identifying features of *translationese*[2] in a corpus of Italian newspaper articles containing translations, (Ilisei et al., 2010; Ilisei and Inkpen, 2011; Pastor et al., 2008) identify document-level trends[3] in Spanish and Romanian translations and comparable original text, with Kurokawa et al. (2009) focusing on a more application-based study of bi-directional machine translation performance using a French−English Hansard corpus, showing that when translation direction is preserved, a smaller corpus in the

[1] The title and concept is a perhaps slightly provocative reference to the 1995 work *The Translator's Invisibility* by the great Lawrence Venuti who posits that the role of a modern translator is to iron out any foreignness in a translated text as not to offend the sensibilities of the Western reader.

[2] The particular stylistic qualities of a translated text, as opposed to one written in the original language. Examples included lower type-token ratio and variations in average sentence length.

[3] Metrics such as average sentence length and type-token ratio which focus on profiling larger trends in written text as opposed to to n-gram features which give an insight into individual word choice.