Found in Translation: to what Extent is Authorial Discriminability Preserved by Translators?

Richard S. Forsyth
16 Halliday Avenue
Leeds LS12 3PQ
U.K.
forsyth_rich [AT] yahoo.co.uk

Phoenix Lam
Department of English Language & Literature
Hong Kong Baptist University
Kowloon Tong
Hong Kong
engplam [AT] hkbu.edu.hk

Corresponding author:
Richard Forsyth
forsyth_rich [AT] yahoo.co.uk

[Please cite as
Forsyth, R.S. & Lam, P.W.Y. (2014). "Found in translation: to what extent is authorial discriminability preserved by translators?" Literary & Linguistic Computing, 29(2), 199-217. doi 10.1093/llc/fqt018
.]

Found in Translation: to what Extent is Authorial Discriminability Preserved by Translators?

Abstract

Most authorship attribution studies have focused on works which are available in the language used by the original author (Holmes, 1994; Juola, 2006) since this provides a direct way of examining an author's linguistic habits. Sometimes, however, questions of authorship arise regarding a work only surviving in translation. One example is *Constance*, the putative "last play" of Oscar Wilde, only existing in a supposed French translation of a lost English original.

The present study aims to take a step towards dealing with cases of this kind by addressing two related questions: (1) to what extent are authorial differences preserved in translation; (2) to what extent does this carry-over depend on the particular translator?

With these aims we analyzed 262 letters written by Vincent van Gogh and by his brother Theo, dated between 1888 and 1890, each available in the original French and in an English translation. We also performed a more intensive investigation of a subset of this corpus, comprising 48 letters, for which two different English translations were obtainable. Using three different indices of discriminability (classification accuracy, Hedge's g, and area under the ROC curve), we found that much of the stylistic discriminability between the two brothers was preserved in the English translations. Subsidiary analyses were used to identify which lexical features were contributing most to inter-author discriminability.

Discrimination between translation sources was possible, though less effective than between authors. We conclude that "handprints" of both author and translator can be found in translated texts, using appropriate techniques.

Keywords: authorship attribution, authorship in translation, inter-textual distance, stylometry, translation 'universals'.

1. Introduction

Using textual characteristics as a guide to the likely author of a document in cases of doubtful or disputed authorship is a field with a long and varied history (Yule, 1944; Mosteller and Wallace, 1964; Wickmannn, 1976; Burrows, 1992; Holmes, 1994; Holmes and Forsyth, 1995; Forsyth et al., 1999; Burrows, 2002; Hoover, 2002; Keselj et al., 2003; Juola, 2006; Grieve, 2007; Koppel et al., 2011). The great majority of such studies concern themselves with analysis of the language in which the texts under consideration were originally written. Sometimes, however, questions of authorship arise regarding a work that only survives in translation. One example is *Constance*, the putative "last play" of Oscar Wilde, only existing in a supposed French translation of a lost English original (de Briel & de Saix, 1954). With rare exceptions, such as Holmes (1992) and Wang and Li (2012), very few studies of authorship in translation have been published. As noted by Hedegaard and Simonsen (2011: 65) "the specific problem of attributing translated texts to their original author has received little attention".

One reason for this lack of attention may be the assumption that the translator's style will somehow dominate, even obliterate, the style of the original author. This assumption is connected with the notion of "translation universals" (Baker, 1993), an idea which postulates that various forms of normalization and homogenization are regularly imposed on translated texts, thus implying that authorial idiosyncrasies will tend to be obscured.

A complicating factor, dubbed by Teich (2003: 145) the "source language shining through", is the finding that translators often bring source-language constructions into the target language, even when they are less than fully natural (Xiao, 2010). This could provide an avenue for elements of authorial style to cross the translation barrier.

Thus the question of the degree to which the original author of a translated text can be recognized remains open, and motivates the twin research questions of the present study: (1) to what extent are authorial differences preserved in translation; (2) to what extent does this carry-over depend on the particular translator? We take a step towards addressing these questions by means of an investigation of the correspondence between the famous impressionist painter Vincent van Gogh (1853-1890) and his brother Theo (1857-1891).

In the next section we give details of the corpus on which our study is based. Section 3 explains our methods of approach, both quantitative and qualitative. Section 4 presents our main results. Finally section 5 outlines the main implications of our findings and points towards some future directions.

2. Corpus Compilation and Anonymization

Two online resources containing letters written by Vincent van Gogh and his brother Theo were used for data collection in the present study. They represent respectively an institutional and an individual effort to chronicle the life of the Dutch painter through his correspondence with family, friends and acquaintances. Both resources are freely accessible on the Internet.

The institutional work, jointly produced by the Van Gogh Museum and Huygens ING, is a collection of 927 letters written and received by Vincent van Gogh from September 1872 to July 1890[1]. A product of 15 years of research, the letters in the collection are annotated with editorial notes and illustrated with relevant artworks. In addition to the provision of the original letters reproduced in facsimile, English translations from Dutch and French originals are also available on the website. Since its publication in 2009 both as a web edition and a book edition, this resource has been considered a definitive collection of van Gogh's letters and has been hailed "the most important art publication" in the press (Thames and Hudson, 2012).

The individual enterprise, created by the author and van Gogh enthusiast David Brooks, and assisted by the translations of H. G. Harrison, forms part of the website *The Vincent van Gogh Gallery*[2]. Representing four years of effort, this resource consists of the English translations of 874 letters written and received by the artist, with a small number of letters between his close relatives, from August 1872 to August 1890. Before the introduction of the van Gogh Museum resource in 2009, the collection at the van Gogh Gallery was one of the rare sources, if not the first, of the painter's collection of letters online.

As our study is concerned with authorial discriminability in the original language as well as in the translated language, both the original letters written by the two brothers and their translations are required for investigation. In this connection, the originals were gathered from the VGM Website. To examine the extent to which authorial discriminability is preserved in different translations of the same original, the English translations from both the VGM and VGG Websites were studied. Since the two resources vary in the number of letters they contain and the composition of their collections,

[1] This resource is referred to as VGM Website thereafter in this article. The website is available at http://vangoghletters.org/vg/.

[2] This resource is referred to as VGG Website thereafter in this article. The website is available at: http://www.vggallery.com/letters/main.htm.

preliminary work has to be done to identify letters which are present in both resources as originals and translations.

After the preliminary filtering stage, three selection criteria were used to extract letters which were suitable for the purpose of the present study. First, only letters which were written from February 1888 onwards were selected. This is because this date marks the final and crucial stage of the painter's life, when he moved from Paris to Arles and embarked on a period of intensive creativity to produce the majority of his most well-known paintings. More importantly, letters from Theo to Vincent before this period are not available, as they were not preserved by the painter, making inter-authorial comparison impossible. Second, only letters which are originally written in French and translated into English were selected. This is because Dutch originals, which are mostly found in correspondence between the two brothers in earlier periods, are small in number. By 1888 the brothers preferred to converse, and correspond, in French. Selecting original letters only in a single language also eliminates the extra variable of source language in authorial discriminability. Finally, to allow for meaningful quantitative textual analysis, only letters which contain more than 80 words were selected. Compared with most previous authorship studies which tend to examine long text blocks containing thousands of words (see, for instance, Holmes, 1998; Craig & Kinney, 2009), the present study represents an attempt to distinguish authors, both in originals and translations, using texts of varying sizes with a relatively short minimum length.

Once the letters were extracted based on the three selection criteria, they were anonymized through the removal of identity-revealing clues. This was achieved through the two means of deletion and replacement. Deletion involves the removal of the address letterhead and superscribed date and/or location in the preamble of the letters. Replacement, on the other hand, involves the substitution of sender and recipient identity by pseudonyms in the salutation and valediction. The body text and the postscript, if present, are kept in their original form for authorial discrimination. Appendix 1 shows an example of a letter before and after the anonymization procedure.

The result, following the selection and anonymization procedures, is a corpus of 291 letters originally written in French and translated into English. More than three quarters (78%) of the letters (N=226) in the corpus are produced by Vincent, with only 36 (12.4%) by Theo. A small number of letters (N=29) written by other authors such as Paul Gauguin (N=16) and Vincent's other family members are irrelevant for the present purpose and are thus excluded from this study. Thus for analytic purposes we are left with a main corpus of 262 letters by the two brothers for which we have both original French and VGM English translations. Basic statistics for this corpus are given in Table 1.

Table 1. Details of Main VGM Corpus.

Author	**Number of Letters**	**Word tokens FR**	**Word tokens EN**
Vincent VG	226	217589	222714
Theo VG	36	19992	20336
Totals =	**262**	**237581**	**243050**

In addition, to investigate translator effects, we also compiled a subset of 48 letters from the main corpus, 35 written by Vincent and 13 by Theo, for which we were able to obtain two English versions. This randomly selected subset includes the 48 French originals and their corresponding English translations, both from the VGM Website, as well as translations from the VGG Website, totalling 144 texts of 133159 words. The median lengths of the anonymized texts in the three sets are respectively 696, 717 and 725 in words. Table 2 shows the composition of this doubly-translated subcorpus.

Table 2. Composition of Doubly-Translated Subcorpus.

Author	**Number of letters**	**Number of word tokens**		
		French originals	**English translations (VGM)**	**English translations (VGG)**
Vincent VG	35	36747	37584	36674
Theo VG	13	7113	7213	7828
Totals =	**48**	**43860**	**44797**	**44502**

3. Methods

Our approach falls within a text-classification paradigm, broadly interpreted, but since this study is an attempt to assess the relative importance of clues left in translated texts by both author and translator, we focus less on discrimination as such than on *discriminability* -- i.e. availability of textual indicators that could be used for discrimination between authors (or between translators).

3.1 Document descriptors

For the purpose of finding features that characterize texts, we take what has become a conventional route, pioneered by Burrows (1992) among others, of using frequently-occurring orthographic word tokens as our basic elements. This means that we characterize differences between texts primarily in terms of differences in lexical choices. This in turn means that our approach does not rely on specialized resources such as parsers, taggers, lemmatizers or content-analytic lexicons, although we report a side study using the LIWC2007 software of Pennebaker & Francis (1996), which taps into some syntactic and semantic characteristics of English texts.

We have written a program in Python3 that takes a corpus and produces a vocabulary of word tokens to be used as a set of textual features. In all experiments reported here the program was given a number N and generated a list of N word tokens in the following manner. Firstly the N most frequent words in each category of the corpus are found and ranked; thus, for example, with letters by Theo and Vincent van Gogh as the two recognized categories, the N most frequent words in the subsamples by each author would be found. Then the two lists are merged by rank-product. Thus a word ranked 10th in one author's list and 5th in the other's would receive a combined rank of 50. (Words absent from either author's top N receive a default rank of N+1 for that author.) Finally the composite word-list is sorted by rank-product and the N items with the smallest rank-products form the vocabulary that will be used.

The idea behind this slight modification of the classic Burrows approach is to ensure that in cases such as the present, where one category dominates the corpus, all categories still have a chance to contribute to the textual features used.

Once a vocabulary has been established, another program converts each text into a numeric feature vector for subsequent processing. This program produces a single line per text with N+4 entries on each line. The N entries are relative frequencies (expressed as percentages) of the words in the chosen vocabulary within the text concerned. The extra four values are the directory/path, filename, category label and size (in words) of that text. None of these four identifying items were used in text-similarity calculations, which only depended on the N vocabulary items.

Choice of the vocabulary size, N, can have an important effect on the results obtained; however, there is no established theory to govern this choice. Burrows (1992; 2002) and colleagues have used values typically ranging from 50 to 150. In the present study, we chose N=69. This is within the range normally used by Burrows and others and allows us to make direct comparisons with results obtained from the LIWC2007 software. In fact, LIWC2007 generates 70 numeric features for each document, but one of these is a word-count which we decided not to use. (Theo and Vincent do differ in the mean lengths of their letters, but this was thought to be an irrelevance to the present study.)

To give an indication of what sort of information is being used to arrive at the results quoted later in this paper, Appendix 2 reproduces the 69 words selected as descriptors from the main corpus (262 letters) both in French and in English.

3.2 Document dissimilarity

The present study is grounded principally on computation of dissimilarities between texts. Given 2 texts represented as vectors of numeric features (e.g. relative word frequencies) the similarity or dissimilarity between them can be calculated in numerous ways. An earlier investigation (Forsyth & Sharoff, 2012) had shown that inverse rank correlation (1-rho, where rho is Spearman's rank correlation coefficient) performed more reliably than many other more commonly used indices (including Cosine distance and Euclidean distance), so for the experiments reported here inverse rank correlation was always used as the dissimilarity measure.

3.3 Evaluation measures

In this paper we use three different measures of document discriminability:

(1) Percentage accuracy of a k-nearest-neighbour classifier;
(2) Hedges' g;
(3) Area under the ROC (Receiver Operating Characteristic) curve.

The most obvious way to measure how well text categories can be distinguished is to apply a classification algorithm to them and record its success rate. Accordingly we implemented a simple but robust classifier, the k-nearest-neighbour classifier (Aha et al, 1991; McKenzie & Forsyth, 1995), and used its percentage accuracy as a discriminability measure. In the results reported here, the value of k (the number of near-neighbours considered) was 5. (It should be stated that this classifier uses a leave-1-out method: the predicted category for any item is the majority category of the k items with the lowest dissimilarity score to the one under consideration -- excluding that item itself, which of course has a dissimilarity of zero.)

However, we are not primarily concerned to find an optimal classification algorithm. Rather we are interested in uncovering potential for discrimination -- i.e. how much evidence to distinguish between authors is preserved in translated texts as compared to original texts. Thus raw classification accuracy, which is known to be susceptible to boundary effects, where error rates can fluctuate quite markedly with slight changes of parameter settings as borderline cases flip from one side of a boundary line to another, is just one index, not necessarily the best, of potential discriminability.

A more sensitive index is Hedges's g. This is an "effect size" measure, proposed by Hedges (1981), one of a number of related measures that seek to quantify the difference between 2 samples. Essentially it expresses the difference between the mean of one sample ($\bar{x}_1$) and the mean of a second sample ($\bar{x}_2$) in terms of the number of standard deviations by which they differ, in effect as a z-score.

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

In contrast to the t-test or z-test the difference between means is divided by an estimate of the pooled standard deviation (s* in the formula below), not the pooled standard error. Thus it is unaffected by the sizes of the 2 groups. This is because it attempts to quantify the average difference between individual members of the 2 groups rather than between the group means themselves.

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

In summary, it provides a standardized index of the size of the difference between 2 sets of scores. In the present experiment we use it to contrast textual dissimilarity scores between texts by the same author with textual dissimilarity scores between texts by different authors.

As a third index of discriminability, we use the area under the ROC (Receiver Operating Characteristic) curve (Bradley, 1997; Fogarty et al., 2005) which is usable in 2-class problems, such as ours. To compute this, we start by defining a function C(x,y) which compares the scores of a positive, and a negative case (x with y), and yields 1 if x exceeds y, 0.5 if they are equal, and 0 otherwise. Then A`, the area under the ROC curve, is defined as

$$A` = \frac{1}{(n \times m)} \sum_{i=1}^{n} \sum_{j=1}^{m} C(x_i, y_j)$$

where n and m are the numbers in each group. A` has an appealing probabilistic interpretation: it equals the probability that a randomly selected positive case will have a higher score than a randomly selected negative case.

This measure is non-parametric: that is to say, it is less affected by outliers or skewed data distributions than, for instance, Hedges's g.

According to Witten and Frank (2005: 173) "no single number is able to capture the tradeoff" between true and false positives in a classification experiment. By using three indices, with differing attributes, we hope to gain a perspective on how well our textual classes can be discriminated.

3.4 Identification of discriminatory lexical features

As supplementary analyses, to assist in interpreting the results concerning discriminability, we performed a keyword analysis (Scott, 2006) and applied the Conditional Inference Tree method (Hothorn et al., 2006) with the aim of identifying which words were contributing most strongly to the discrimination between authors.

4. Results

4.1 Discriminability results using the full corpus

We applied the methodology described in section 3 to the main corpus of letters by Theo and Vincent van Gogh, in French and English. The advantage of this dataset is that it is large (262 letters comprising 237,581 word tokens, in French). A disadvantage is that it is lopsided: 226 texts are by Vincent, representing over 86% of the letters (over 91% in terms of total words). A further disadvantage is that we only have one translation source, VGM, for these texts.

Table 3 shows the scores on our three indices of discriminability for the French originals and the English translations.

Table 3. Discriminability of original and translated letters (VGM).

	5-Nearest-Neighbour success rate (%)	**Hedges's g**	**Area under ROC curve (A`)**
French originals (n=262)	95.04	1.4000	0.8469
English translations (n=262)	90.46	1.0036	0.7637
Ratio (English/French):	0.9518	0.7169	0.9018

In this table Hedges's g and A` have been computed between two groups of calculated dissimilarities, namely same-author inter-text dissimilarities versus different-author inter-text dissimilarities. In other words, the dissimiliarity between every text and every other text in the corpus was computed, as described in subsection 3.2, and these values were placed in two groups depending on whether the two texts were by the same author or not; then g and A` were computed using these two groups of scores.

The last row of Table 3 gives the ratio of the preceding two rows (En/Fr). These ratios can be regarded as estimates of how much discriminability remains after translation. Clearly there is some degradation in each column: the level of inter-author discriminability in the translated texts is somewhere between 72% and 95% of that of the originals. In fact, the nearest-neighbour classifier makes nearly twice as many mistakes on the English letters as on the French (25 versus 13). Nevertheless, a success rate of over 90% on the translated texts is a respectable performance. It indicates that a substantial fraction of whatever authorial habits of lexical preference enable the classifier to distinguish between the two brothers have been carried over into the English translations.

This impression is reinforced by considering the ratio of the A` values (0.9018). This -- the median of our three ratios -- can be construed as indicating that the separability of the 2 authors in English translation is more than 90% of their separability in the original language.

An illustration of the level of discriminability between the two authors in French and in English translation can be seen in Figures 1 and 2. Figure 1 summarizes 34191 inter-text dissimilarities; that is, the dissimilarities between every letter and every other letter in the main corpus (VGM French originals, N=262). These have been placed in three groups: (1) both texts by Vincent, (2) each text by a different author, and (3) both texts by Theo. The horizontal line within each box indicates the median level of dissimilarity in that group. The boxes enclose the values between the upper and lower quartiles. Figure 2 plots the same information for the 34191 inter-text dissimilarities among the 262 letters of the VGM English translations.

[Figure 1. Inter-text dissimilarities within authors and between authors, French originals.]

It can be seen that in both cases Vincent is more distinctive than Theo. That is to say, there is less overlap between the inter-text dissimilarities among Vincent's letters and the cross-author dissimilarities than there is between the inter-text dissimilarities among Theo's letters and the cross-author dissimilarities. This tends to support the assertion "lines from the letters chosen at random immediately reveal his pen and his spirit" -- referring to Vincent -- which can be found on the VGM website[3].

[Figure 2. Inter-text dissimilarities with authors and between authors, VGM English translations.]

In addition, comparison of Figure 2 with Figure 1 shows that while this asymmetry is preserved in translation, the separation between within-author and between-author dissimilarities is weaker in

[3] found at: http://vangoghletters.org/vg/letter_writer_4.html#intro.I.4.6

English than in the French originals, thus providing a visual confirmation of the numerical results in Table 3. Nevertheless, separability is far from obliterated.

4.2 Results on a multiply-translated subset of the corpus

With the main corpus of 262 letters, we only have one translation source, the Van Gogh Museum (VGM). In order to make an initial appraisal of whether this level of translator effect is ordinary or exceptional, we also performed several analyses on a smaller subset of the letters for which it was possible to obtain alternative translations.

As shown in section 2, this sub-corpus consists of 48 letters -- 13 written by Theo and 35 by Vincent. It comprises 43,860 word tokens in French. It was originally compiled as a random selection of 42 letters by the 2 brothers from the main corpus, but with only nine letters by Theo we considered it unreasonably vulnerable to small-sample effects and thus decided to augment it with a slight oversampling from the less well represented category. Consequently we added four more letters, randomly selected from those by Theo that were not already present as well as 2 more letters randomly selected from those by Vincent that were not already present. This subcorpus is still relatively unbalanced, with nearly 73% of the texts by one author (Vincent) and thus represents quite a challenging classification problem.

The main point, however, is that it enables us to compare translated versions from different sources, since we have English translations for all these letters both from the Van Gogh Museum (VGM) and the Van Gogh Gallery (VGG). In addition, just to establish a baseline, we also translated each of these 48 letters from French into (traditional) Chinese, using the online Google-translate service (http://translate.google.com/). We intended this to be a "zero-level" baseline, since French and Chinese are from different language families and the automatic translation service does not claim to come close to professional quality. However, if authorial distinction could survive passage through such a "noisy channel" it would open up a number of previously unsuspected practical applications.

Table 4 shows the values of our three discriminability indices for four versions of this subcorpus: the French originals, the VGM translations into English, the VGG translations into English, and the Google translations into Chinese.

Table 4. Discriminability indices on the multiply-translated subcorpus,

Subcorpus	**5-Nearest-Neighbour success rate (%)**	**Hedges's g**	**Area under ROC curve (A`)**	**A` ratio (translation / original)**
Original French	95.83	1.3145	0.8273	1.00
Museum English	85.42	0.8444	0.7243	0.88
Gallery English	83.33	0.8302	0.7173	0.87
Google Chinese	79.17	0.4886	0.6323	0.76

The nearest-neighbour success rate shows a slightly greater degradation from French to English with this subset than with the main corpus of 262 letters. It also suggests that the two translation sources differ negligibly in this effect. However, we place less weight on classifier success rate than the other measures because it is based, in effect, on a smaller sample. In this case, it arises from 48 decisions, whereas the other two measures result from 1128 inter-text comparisons (595 comparisons of every letter by Vincent with every other letter by Vincent, 78 comparisons of every letter by Theo with every other letter by Theo, plus 455 comparisons of every letter by Vincent with every letter by Theo).

According to Hedges's g and A` the Gallery translations are scarcely worse in terms of potential discriminability than the Museum translations. The final column of this table uses A` as probably the most stable of our indices -- giving the ratio of A` in the translated case to the original. Roughly speaking, 87% or 88% of the original authorial discriminability is preserved in the English translations, whereas about only 76% is preserved in the Google Chinese translations.

Another way of assessing the impact of translation is to perform a similar analysis on the translated texts, but this time, instead of trying to distinguish the original authors, we attempt to distinguish texts from one source (VGM) from those of the other (VGG).

Table 5 shows values of our three discriminability indices for the combined English subcorpora (96 letters) when the goal was to distinguish Museum translations from Gallery translations.

Table 5. Discriminabilty of translation sources.

	5-Nearest-Neighbour success rate (%)	**Hedges's g**	**Area under ROC curve (A`)**
English Translations (n=96)	77.08	0.5678	0.6652

Here the classifier's success rate is substantially better than chance, though the other two measures indicate that the degree of differentiation between the two translation sources is rather less than that achieved between the two authors. Comparison of Tables 4 and 5 suggests that in this case it is somewhat more difficult to distinguish translators than original authors (in translation).

4.3 Analysis using LIWC (Linguistic Inquiry and Word Count)

As a side-study, we applied the LIWC2007 program (Pennebaker & Francis, 1996) to both our English translations of the subcorpus of 48 letters. LIWC2007 is a program that computes, for any text in English, 70 numeric variables (one of which is a word-count, which we ignore). The other 69 are a mixture of textual, lexical, syntactic and content-based variables, such as the examples below:

Qmarks	percentage of sentences ending with question marks;
ppron	percentage of personal pronouns;
adverb	percentage of adverbs;
negemo	percentage of negative emotional words;
body	percentage of words dealing with body parts or bodily functions;
money	percentage of words related to money or finance.

The reasons for investigating discriminability using the LIWC2007 variables, even though the software is limited to English-language texts, were twofold. In the first place, it uses a wider variety of linguistic features than merely individual word occurrence rates and therefore provides a benchmark against which to assess whether the frequent-word approach is needlessly ignoring useful information. Secondly, it serves quite well as a model for a realistic application scenario.

To explicate the second point we can ask: why would anyone want to assign an author to a text known to be translated from an original language? One answer might be for the sake of "pure" scholarship. We could envisage a scenario in which a Latin manuscript is discovered that purports to be a translation of a lost Platonic dialogue. In such a case scholars world-wide would feel a pressing need to make an informed judgement regarding its likely authenticity (by comparison not with Greek originals, which would be impossible, but with other Latin translations of genuine works by Plato).

Another plausible motivation would be a case in which, for example, a series of email messages is received which are in English but known to have been translated from another language, e.g. Arabic or Chechen, before being sent. It may be important to establish whether at a particular point in the series the person composing them changed. Alternatively, it may simply be more convenient to work in a major language like English or Chinese than in an under-resourced language like Estonian or Chuvash because more analytic software is available.

Table 6. Discriminability of translated texts using LIWC features.

Results using LIWC2007 variables	**5-Nearest-Neighbour success rate (%)**	**Hedges's g**	**Area under ROC curve (A`)**	**A` ratio (translation / original)**
Museum English	79.17	0.3746	0.6164	0.75
Gallery English	81.25	0.4036	0.6343	0.77

Table 6 gives the results on the two different English translations of our 48-item subset of the letters. On all three discriminability measures the LIWC results are inferior to the corresponding frequent-word-based results. However, the possibility remains open that LIWC, by tapping into somewhat different aspects of the linguistic information contained in a text, could be used to augment the performance of a purely word-based system. (This also applies, of course, to alternative text-processing systems.)

4.4 Using Conditional Inference Trees to highlight discriminatory features

So far we have looked only at aggregate overall discriminability scores, but an analyst would also very probably be interested in gaining some insight into which linguistic features are most important in differentiating the textual categories.

To throw some light on which features are most useful in discriminating between our two authors, in French or in translation, we took advantage of the "Conditional Inference Tree" package (Hothorn et al., 2006) in the R programming environment (R Development Core Team, 2012). This uses a recursive partitioning strategy to generate a tree-structured collection of tests that can be used to sort a dataset into categories -- each test being a comparison of a variable with a constant (in our case a percentage rate with a threshold level). For all trees we used the 48-item subsample as data and set the maximum tree depth to 4, though in fact the system never generated a tree with depth greater than 3.

[Figure 3. Discrimination tree for Theo versus Vincent, French originals.]

Figure 3 shows the tree generated from the VGM French data. The root node, i.e. most important test, uses the feature *et*. A low rate indicates a letter by Theo, a high rate leads to a split on the variable *ton*: a low rate of *ton* leads to a leaf node of 29 letters by Vincent, a higher rate leads to a mixed leaf node, with 1 letter by Theo and 6 by Vincent.

As will be discussed in subsection 4.5, the word "*et"* is not a Vincent marker in its own right. Rather its usefulness as a discriminator arises from Theo's tendency to use the ampersand ("*&*") much more frequently than Vincent -- a tendency, incidentally, that is masked in English, where the translators render both "*et"* and *&* as "*and"*.

[Figure 4. Discrimination tree for Theo versus Vincent, VGM English translations.]

Figure 4 shows the tree generated from the VGM English translations. Here the root node uses the word "*your*" and there is no further subdivision. This divides the data into 35 letters with a low rate of "*your*" (3 by Theo) and 13 letters with a high rate of "*your*" (10 by Theo). If we term the items in the minority class at any leaf node as "exceptions", this gives 6 exceptions out of 48 or 12.5%.

Rather than show a multiplicity of different trees, we summarize their main properties in Table 7. This gives, for each dataset derived from the multiply-translated subcorpus of 48 letters, the variable chosen as the root node, any lower-level variables selected, and the number of "exceptions" in the tree.

Table 7. Summary of tree-growing experiments.

Dataset (n=48)	**Root variable**	**lower-level variables**	**"Exceptions"**
VGM French	"et"	"ton"	1
VGM English	"your"	-	6
VGG English	"you"	"your"	5
Google Chinese	"你" [utf8 code = 4f60, nǐ = "you"]	-	0
VGM English, LIWC variables	you [all second person pronouns, including you, your & yours]	-	4
VGG English, LIWC variables	you [all second person pronouns, including you, your & yours]	conj [conjunctions]	3

What is striking is the presence of variables related to usage of the second-person pronoun in all trees. In all the translations a second-person marker appears in the root test. Even in the French data, where *et* acts as a kind of shadow of Theo's fondness for the ampersand, *ton* appears at level 2.

We defer discussion of the light this may throw on the relationship between the two brothers till subsection 4.5, but note here that a tree based on the Google Chinese translations, in effect, makes no mistakes, simply by using the character that normally translates "*you*", despite the fact that the 5NNC results on the Google translations were comparatively poor. Thus the information necessary for perfect discrimination between the two authors is present in the Chinese translations, even though it was not fully exploited by the nearest-neighbour classifier. This would seem to be a case where adding 68 features degrades the performance of a single highly informative marker.

[Figure 5. Discrimination tree for Museum versus Gallery translations.]

Figure 5 shows the tree grown with the objective of discriminating the two English translation sources, VGM and VGG. This uses "*i'm*" as root node (displayed as "i_m" to avoid problems within the R system) and "*it's*" and "*am*" as second-level tests. Using "exceptions" as defined above, this tree contains 8 exceptions out of 96 (8.33%). This tree reveals a greater tendency of the VGM translators to use contracted forms such as "*I'm*", "*I've*" and "*it's*" than the VGG translator, who is more likely to employ "*I am*", "*I have*" and "*it is*".

Taken together, these analyses suggest that information that would allow accurate identification both of the author and of the translator is present in the translated letters, though employment of more than a single approach may well be necessary to uncover it.

4.5 Keyword-oriented analyses

VGM French originals:
According to the statistics generated by WordSmith (Scott, 2006), letters by Vincent are generally longer (1117 words per letter) than Theo’s (581 words per letter). Apart from writing longer letters, Vincent also preferred longer sentences, as shown by the higher mean sentence length (24 words per sentence) when compared with his brother (19 words per sentence). As regards the standardised type-token ratio, the figure generated from letters by Vincent is slightly higher (42.65) than that from letters by Theo (40.20), possibly suggesting a wider range of words used by the painter.

For a keyness-based analysis, the 48 letters of the doubly-translated subcorpus were used, with Theo’s letters as the focus corpus and Vincent’s letters as the reference corpus. In this analysis (unlike those of the preceding sections) punctuation symbols as well as word tokens were counted. The measure of keyness used was the log-likelihood version of Chi-squared (Dunning, 1993), and all tokens with a keyness score greater than 25.26 were listed. This threshold corresponds to a p-value of less than 0.0000005, although such probabilities need not be taken at face value since this is not a hypothesis-testing context. Table 8 lists the 10 tokens with a keyness score exceeding 25.26, plus or minus. (In this, as well as the next two tables, "~0" is used to indicate a p-value less than 0.000000005.)

Table 8. Key tokens in Theo's letters compared with Vincent's (VGM French originals).

N	Item	Freq. (Theo)	% (Theo)	Freq. (Vinc)	% (Vinc)	Keyness value	p-value
(1)	&	130	1.65	61	0.15	253.42	~0
2	tu	139	1.76	253	0.63	84.04	~0
3	il	145	1.84	347	0.86	52.50	~0
4	iҫi	12	0.15	0	0.00	43.47	~0
5	tes	19	0.24	6	0.01	43.42	~0
6	chez	31	0.39	35	0.09	33.54	0.00000001
7	voir	33	0.42	45	0.11	29.33	0.00000006
8	ton	26	0.33	32	0.08	25.82	0.00000037
-1	et	33	0.42	1041	2.58	-196.28	~0
(-2)	--	3	0.04	283	0.70	-78.57	~0

Table 8 shows the seven positive keywords and one negative keyword in Theo's letters when compared with Vincent's, along with one positive and one negative punctuation symbol. Specifically, the seven positive keywords can be classified into three groups: pronouns (*tu, il, tes, ton*), location-oriented prepositions and adverbs (*iҫi, chez*), and a verb (*voir*).

The personal pronouns *tu*, *il* and the possessive determiners *tes, ton* are significantly more frequently used by Theo than Vincent. The proportionally more frequent use of the second person pronoun *tu* and the second person possessive determiner *tes* by Theo possibly suggests that the younger brother was more recipient-oriented. The third person singular masculine pronoun *il*, on the other hand, indicates that Theo was also interested in writing about others (e.g. *Mr Lauzet*, *Mr Peyron*) apart from directly addressing his elder brother. The fact that these pronouns are significantly less frequent in Vincent's letters may imply that the painter was more self-absorbed and less interested in his addressee and other people surrounding his life.

Regarding the use of location-oriented prepositions and adverbs, there is a significantly more frequent use of *chez* in Theo's letters than Vincent's. In such instances, *chez* are commonly followed by *nous, lui* and proper names (e.g. *Bernard*, *Tangui*), suggesting aspects of family and social life with his wife and friends such as dining and gathering. When *chez* was used by Vincent, in contrast, it was more frequently followed by *toi* or *moi*. Since the two brothers were physically remote from each other and visits were difficult, the collocates of *chez* in Vincent's letters seem to indicate a more solitary life.

It is not entirely clear why Theo uses the verb infinitive *voir* at a significantly higher rate than Vincent. He also uses the participle forms (*voyant*, *vu*) relatively more frequently than Vincent (14/7113 or 0.1986% versus 35/36747 or 0.0952%) but the conjugated tenses of this verb (present, imperfect, future, conditional and simple past) are used at almost the same rate by both brothers (22/7113 versus 97/36747).

This infinitival imbalance may be an indirect reflection of Theo's role as a picture dealer, whose daily work involved arranging for others to view images; whereas Vincent's vocation demanded that he look at things himself. An inspection of the 78 concordance lines containing *voir* (33 from Theo's, 45 from Vincent's letters) lends some credence to this conjecture. In Theo's case the person or persons doing the seeing is most commonly a third party or parties other than either brother (21/33) whereas in Vincent's letters the most common seer is Vincent himself (19/45). As for the person(s) or object(s) being seen, with Theo it is a person or persons 16 times out of 33, a picture or pictures 10 times and some other inanimate entity 7 times. In Vincent's letters the thing seen, visually or metaphorically, is a person or persons 7 times out of 45, a picture or pictures 12 times and another object or state of affairs 26 times. With Theo, the two equally most frequent combinations of seer and seen are someone else seeing a picture (6/33 cases) and someone else (such as Mr Salles) seeing Vincent (also 6/33). For Vincent the most common combination is Vincent himself seeing something other than a picture (13/45), often a state of affairs rather than a physical entity. The concordance lines below, the first two by Theo the next two by Vincent, illustrate these differences.

```
838f.txt  est gentil de mr salles d'être allé te voir . je lui avais écrit au jour de l'an
888f.txt  à montrer . lauzet est venu hier matin voir tes tableaux , il est très occupé avec
630f.txt  . car cela me ferait grand plaisir de voir qu'ils aient trouvé une voie . poignée de
785f.txt  n'ai ni femme ni enfant j'ai besoin de voir les champs de blé et difficilement je
```

When compared with all the seven positive keywords, the only negative keyword *et* has the highest keyness value. This conjunction was used proportionally much more frequently by Vincent than Theo. A close examination of the letters shows that instead of using *et* for coordination, Theo preferred the symbol ampersand (*&*), the only positive punctuation marker, making it potentially an important stylistic marker for authorial discrimination between the two brothers.

Finally, the only negatively scoring punctuation symbol in this list ("--") is much more common in Vincent's letters than those by Theo. This may well be an indicator of Vincent's more informal -- and parenthetical -- writing style.

VGM English translations:
Similar to the French originals, the English translations of letters by Vincent are longer (1114 words per letter) than Theo's (580 words per letter). The mean number of words per sentence is also higher for Vincent (23 words per sentence) than for Theo (19 words per sentence). However, the standardised type-token ratios of the translated letters produced by the two brothers are similar (40.74 for Vincent and 40.50 for Theo).

In terms of key tokens, results were generated using the VGM English translations of Theo's letters as the focus corpus and Vincent's translated letters as the reference corpus. Four words are found to have keyness values greater than 25.26, using the log likelihood measure (Table 9).

Table 9. Key tokens in the VGM English translations of Theo's letters when compared with those of Vincent's letters.

N	Item	Freq. (Theo)	% (Theo)	Freq. (Vinc)	% (Vinc)	Keyness value	p-value
1	he	99	1.23	198	0.47	53.93	~0
2	your	65	0.81	103	0.24	50.01	~0
3	you	176	2.18	515	1.22	40.60	~0
4	her	22	0.27	18	0.04	31.90	0.00000002
(-1)	--	3	0.04	273	0.64	-73.00	~0

Table 9 shows the four positive keywords in the English translations of Theo's letters when compared with those of Vincent's. No negative keyword is found. The four positive keywords are similar to those found in the French originals in the sense that they are pronouns and/or determiners (*he*, *you*, *your*, *her*). Unlike the French originals, however, other word types, such as prepositions and adverbs are not found to be significantly more frequent. Further, the third person singular feminine pronoun and determiner (*her*) is a positive keyword in the English translations but not the French originals. The frequent occurrences of these second and third personal pronouns and possessive determiner in the English translations of the letters are consistent with the finding from the French originals, meaning that certain aspects of discriminability between the two authors is still preserved in translation.

The only negative item is the dash ("--"), again more frequent in Vincent's letters than Theo's.

VGG English translations:
Table 10 shows the five words and one punctuation symbol (semi-colon) with the highest keyness scores in the VGG translations. It can be seen that these are very similar to those from the VGM translations, once again highlighting the relative prevalence of second-person and third-person pronouns in Theo's writings. The single negative item ("--") again indicates Vincent's fondness for dashes.

Table 10. Key tokens in the VGG English translations of Theo's letters when compared with those of Vincent's letters.

N	Item	Freq. (Theo)	% (Theo)	Freq. (Vinc)	% (Vinc)	Keyness value	p-value
1	your	78	0.89	101	0.25	65.71	~0
2	you	221	2.53	563	1.37	54.70	~0
3	he	127	1.46	263	0.64	51.84	~0
(4)	;	47	0.54	77	0.19	28.96	0.00000007
5	her	22	0.25	18	0.04	28.61	0.00000009
6	she	25	0.29	26	0.06	26.51	0.00000026
(-1)	--	11	0.13	214	0.52	-32.75	0.00000001

Overview:
Taken together, Tables 8, 9 and 10 gives us some idea of what authorial indicators are preserved and which are being lost in the process of translations.

Firstly we see that, at a given p-level, there are more authorially distinctive words in the original French than in the English translations, which helps to explain the loss of discriminability in translation.

Secondly we see that both translation sources have a very similar effect, compatible with our earlier finding that neither translation is markedly better or worse in terms of author discriminability.

Thirdly we find that all three versions of the letters give evidence of difference in orientation between the two brothers by highlighting Theo's references to his reader and to third persons. This fits well with the idea of Theo as an outwardly-directed supporter of Vincent the inspired loner, lost in his personal artistic vision.

When we look at which keywords do get "lost in translation", we find some support for the notion that *normalization* is a translation universal. Two of Theo's notable quirks, his predilection for the ampersand and his persistent misspelling of *ici* as *içi* (with an unnecessary cedilla) are silently smoothed away in translation. Thus the translated text is rendered more standard or normal than the original. This underlines the importance of only using words, omitting punctuation, in the discrimination experiments, since foibles of punctuation and spelling tend to be silently emended in the translation process.

6. Discussion

The present investigation can only be regarded as a "ranging shot". However, owing to the relative dearth of previous work on the questions examined here, a single case study such as the present can serve as a valuable starting point in assessing the relative contributions of original author on the one hand and translator on the other.

Our results suggest that, as expected, some signs of authorship are lost in translation; but many are preserved. Clearly it is wrong to assume that all evidence of the original author is obliterated by translation. We also observed translator effects. Thus this pilot study leads to a working hypothesis that both the author's and the translator's "handprints" are present in a translated work and, with suitable tools, both can be revealed. Future work on a wider variety of authors and translators will enable us to give more precise account of the relative importance of author and translator effects in translated works, as well as guidance on how to find them.

From a practical point of view, the fact that seeking authorial indicators in translated texts is not futile opens up a number of potential applications. In today's world of global inter-connectedness, many texts are translations. Establishing that authorship attribution in translated texts is no longer a no-go area considerably extends the range of attribution problems that can, in principle, be tackled. Such problems include forensic applications as well as literary ones, including harassing emails, plagiarized articles, ransom demands and terrorist announcements.

References

Aha, D.W., Kibler, D. & Albert, M.K. (1991). Instance-based learning algorithms. *Machine Learning*, 6, 37-66.

Baker, M. (1993). “Corpus Linguistics and Translation Studies: Implications and Applications.” In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis and E. Tognini-Bonelli, eds., 233-250. Amsterdam: Benjamins.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30, 1145-1159.

Burrows, J.F. (1992). Not unless you ask nicely: the interpretive nexus between analysis and information. *Literary & Linguistic Computing*, 7(2), 91-109.

Burrows, J.F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary & Linguistic Computing*, 17(3), 267-287.

Craig, H. & Kinney, A.F. (2009) eds. *Shakespeare, Computers and the Mystery of Authorship*. Cambridge: C.U.P.

De Briel, H. & de Saix, L.G. (1954). *Constance: comédie en quatre actes*. Fayard.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61-74

Fogarty, J., Baker, R., Hudson, S. (2005) Case Studies in the use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction. *Proceedings of Graphics Interface (GI 2005)*, Waterloo, Canada, May 2005, 129-136.

Forsyth, R.S., Holmes, D.I. & Tse, E. (1999). Cicero, Sigonio and Burrows: investigating the authenticity of the 'Consolatio'. *Literary & Linguistic Computing*, 14(3), 375-400.

Forsyth, R.S. & Sharoff, S. (2012). Quantifying document dissimilarity within and across languages: a benchmarking trial. *6th Inter-Varietal Corpus Studies (IVACS)* Conference, Leeds, 21-22 June 2012.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary & Linguistic Computing*, 22(3), 251-270.

Hedegaard, S. & Simonsen, J.G. (2011). Lost in translation: authorship attribution using frame semantics. Proc. *49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, June 2011, 65-70.

Hedges, L.V. (1981). "Distribution theory for Glass's estimator of effect size and related estimators". *Journal of Educational Statistics* **6** (2): 107–128. doi:10.3102/10769986006002107.

Holmes, D.I. (1992). A stylometric analysis of Mormon scripture and related texts. *J. Royal Statistical Society (A)*, 155(1), 91-120.

Holmes, D.I. (1994). Authorship attribution. *Computers & the Humanities*, 28, 1-20.

Holmes, D.I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), 111-117.

Holmes, D. I. and Forsyth, R. S. (1995). The *Federalist* revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2), 111-127.

Hoover, D.L. (2002). Frequent word sequences and statistical stylistics. *Literary & Linguistic Computing*, 17(2), 157-180.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651--674.

Juola, P. (2006). Authorship attribution. *Foundations & Trends in Information Retrieval*, 1(3), 233-334.

Keselj, V., Peng, F., Cercone, N. & Thomas, C. (2003). N-gram based author profiles for authorship attribution. *Proc. Pacific Association for Computational Linguistics*, Dalhousie University, Nova Scotia, August 2003, 255-264.

Koppel, M., Schler, J. & Argamon, S. (2011). Authorship attribution in the wild. *Language Resources & Evaluation*, 45, 83-94.

McKenzie, D.P. & Forsyth, R.S. (1995). Classification by Similarity: An Overview of Statistical Methods of Case-Based Reasoning. *Computers in Human Behavior*, 11(2), 273-288.

Mosteller, F. & Wallace, D.L. (1964). *Inference and Disputed Authorship: the Federalist*. Reading, Massachusets: Addison-Wesley.

Pennebaker, J.W., & Francis, M.E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion, 10*, 601-626.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Scott, M. (2006). The Importance of Key Words for LSP. In Arnó Macià, E., A. Soler Cervera & C. Rueda Ramos (eds.), *Information Technology in Languages for Specific Purposes: issues and prospects*. New York: Springer, pp. 231-243.

Teich, E. (2003). *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

Thames & Hudson. (2012). Vincent van Gogh -- The Letters. Available at http://www.thamesandhudson.com/9780500238653.html. Accessed 27 July 2012.

Wang, Q. & Li, D. (2012). Looking for translator's fingerprints: a corpus-based study on Chinese translations of Ulysses. *Literary & Linguistic Computing*, 27(1), 81-93.

Wickmann, D. (1976). On disputed authorship, statistically. *ALLC Bulletin*, 4(1), 32-41.

Witten, I.H. & Frank, E. (2005). *Data Mining*, 2nd edition. San Francisco: Morgan Kaufmann.

Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, 15(1), 5-35.

Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary.* Cambridge: C.U.P.

Appendix 1. Example of a letter before and after the anonymisation procedure. The different parts are marked in bold.

Before anonymization	
Goupil & C° **imprimeurs éditeurs** **Boussod, Valadon & Cie** **successeurs** **19, Boulevard Montmartre, Paris** **Adr · Télégr · Boussoval · Paris.** **le 29 Juillet 1889**	*Preamble* delete address letterhead, superscribed date and/or location
Mon cher **Vincent**,	*Salutation* replace recipient identification
Je suis un peu inquiet que tu n'aies pas reçu ma lettre qui contenait un bon de fr 10.-- Généralement tu écris de suite après réception, sans cela je dirais que tu n'aies pas eu le temps. Je me reproche que je t'écris si rarement, mais écrire des lettres m'est dans le dernier temps extrêmement difficile, je ne sais pas à quoi cela tient. J'ai reçu en parfait état ton dernier envoi que je trouve extrêmement beau. Sont ce des choses que tu avais mis de côté exprès pour les laisser sécher, car je trouve dans la pluspart de ces toiles plus de clarté d'expression et un si bel ensemble. Le sous bois avec les arbres entourés de lierre, la promenade à Arles & les champs avec les jardins au printemps sont bien beaux, ceux çi et encore d'autres sont montés maintenant sur des chassis dont nous avons enlevé ceux qui s'y trouvaient et sont actuellement chez Tangui. Ils font très bien dans le cadre. Tangui lui-même les aime beaucoup aussi. Je trouve que tu choisis de beaux sujets pour tableaux, ces arbres touffus pleins de fraîcheur et baignés dans la lumière du soleil sont merveilleusement beaux. Si tu vivais dans un entourage entièrement à ton gout et que tu étais entourés de gens que tu aimais & qui te rendaient ton amitié, je serais très content car tu ne peux pas mieux travailler que tu le fais. et quelle quantité de belles choses n'as tu pas produites. Il est heureux que ta santé va bien. Mr Peyron m'écrivait dernièrement qu'il trouvait ton état très satisfaisant. Espérons que cela ira de mieux en mieux. Les parents de Jo sont içi en ce moment, la mère chez nous & le père chez André. Surtout pour Jo c'est une bonne distraction et cela la force à prendre du mouvement ce qui parait être nécessaire. Elle a bonne mine & est seulement un peu faible. Moi j'ai l'air d'un cadavre, mais j'ai été voir Rivet qui m'a donné toute espèces de drogues qui ont cependant cela de bon qu'elles ont fait cesser ce toux qui me tuait. Je crois que cela est passé maintenant. C'est le changement de vie et avec la façon dont je suis soigné maintenant je vais, une fois le mal passé, reprendre des forces. Hier nous avons été tous à St Germain. oh que la campagne est pourtant belle. Pourquoi les gens vont ils s'esquinter dans les villes quand ils pourraient respirer de ce bon air qui redonne la vie. Est ce que tu sors quelquefois maintenant de l'établissement? Ecris moi quand tu peux -- dis moi un peu comment cela va, ne travaille pas trop. Bonne poignée de mains, aussi de Jo.	*Body* keep its original form

à toi **Theo** Merci mille fois du bel envoi.	*Valediction* replace sender identification *Postscript* leave alone if present

After anonymization

Mon cher **Lector1**,

Je suis un peu inquiet que tu n'aies pas reçu ma lettre qui contenait un bon de fr 10.-- Généralement tu écris de suite après réception, sans cela je dirais que tu n'aies pas eu le temps. Je me reproche que je t'écris si rarement, mais écrire des lettres m'est dans le dernier temps extrêmement difficile, je ne sais pas à quoi cela tient. J'ai reçu en parfait état ton dernier envoi que je trouve extrêmement beau. Sont ce des choses que tu avais mis de côté exprès pour les laisser sécher, car je trouve dans la pluspart de ces toiles plus de clarté d'expression et un si bel ensemble. Le sous bois avec les arbres entourés de lierre, la promenade à Arles & les champs avec les jardins au printemps sont bien beaux, ceux çi et encore d'autres sont montés maintenant sur des chassis dont nous avons enlevé ceux qui s'y trouvaient et sont actuellement chez Tangui. Ils font très bien dans le cadre. Tangui lui-même les aime beaucoup aussi. Je trouve que tu choisis de beaux sujets pour tableaux, ces arbres touffus pleins de fraîcheur et baignés dans la lumière du soleil sont merveilleusement beaux. Si tu vivais dans un entourage entièrement à ton gout et que tu étais entourés de gens que tu aimais & qui te rendaient ton amitié, je serais très content car tu ne peux pas mieux travailler que tu le fais. et quelle quantité de belles choses n'as tu pas produites. Il est heureux que ta santé va bien. Mr Peyron m'écrivait dernièrement qu'il trouvait ton état très satisfaisant. Espérons que cela ira de mieux en mieux.

Les parents de Jo sont içi en ce moment, la mère chez nous & le père chez André. Surtout pour Jo c'est une bonne distraction et cela la force à prendre du mouvement ce qui parait être nécessaire. Elle a bonne mine & est seulement un peu faible. Moi j'ai l'air d'un cadavre, mais j'ai été voir Rivet qui m'a donné toute espèces de drogues qui ont cependant cela de bon qu'elles ont fait cesser ce toux qui me tuait. Je crois que cela est passé maintenant. C'est le changement de vie et avec la façon dont je suis soigné maintenant je vais, une fois le mal passé, reprendre des forces. Hier nous avons été tous à St Germain. oh que la campagne est pourtant belle. Pourquoi les gens vont ils s'esquinter dans les villes quand ils pourraient respirer de ce bon air qui redonne la vie. Est ce que tu sors quelquefois maintenant de l'établissement?

Ecris moi quand tu peux -- dis moi un peu comment cela va, ne travaille pas trop. Bonne poignée de mains, aussi de Jo.

à toi
Prae

Merci mille fois du bel envoi.

Appendix 2 -- Words selected by the feature-finding program from the main corpus (262 letters)

VGM French originals:

de	que	je	à	le	la	il	et	tu	pas	les	un
bien	ce	pour	en	ne	des	est	a	dans	une	qui	
mais	nous	si	cela	plus	comme	y	me	te	c'est	avec	qu'il
du	lui	tout	j'ai	au	fait	mon	encore	faire	moi	aussi	très
se	toi	voir	par	beaucoup		ou	peu	chez	même	on	sont
là	jo	ta	lettre	sur	temps	ici	ton	ces	tes	cette	

VGM English translations:

the	to	and	of	that	you	a	i	in	it	for	
have	but	is	as	be	he	me	with	very	if	at	one
your	on	my	it's	from	which	are	this	so	all	there	like
do	good	him	what	see	will	his	was	i'm	more	would	by
we	much	has	well	or	not	about	them	here	when	who	can
little	an	i've	come	than	had	now	jo	don't	letter		

===========

List of Tables:

List of Figures:

Figures.

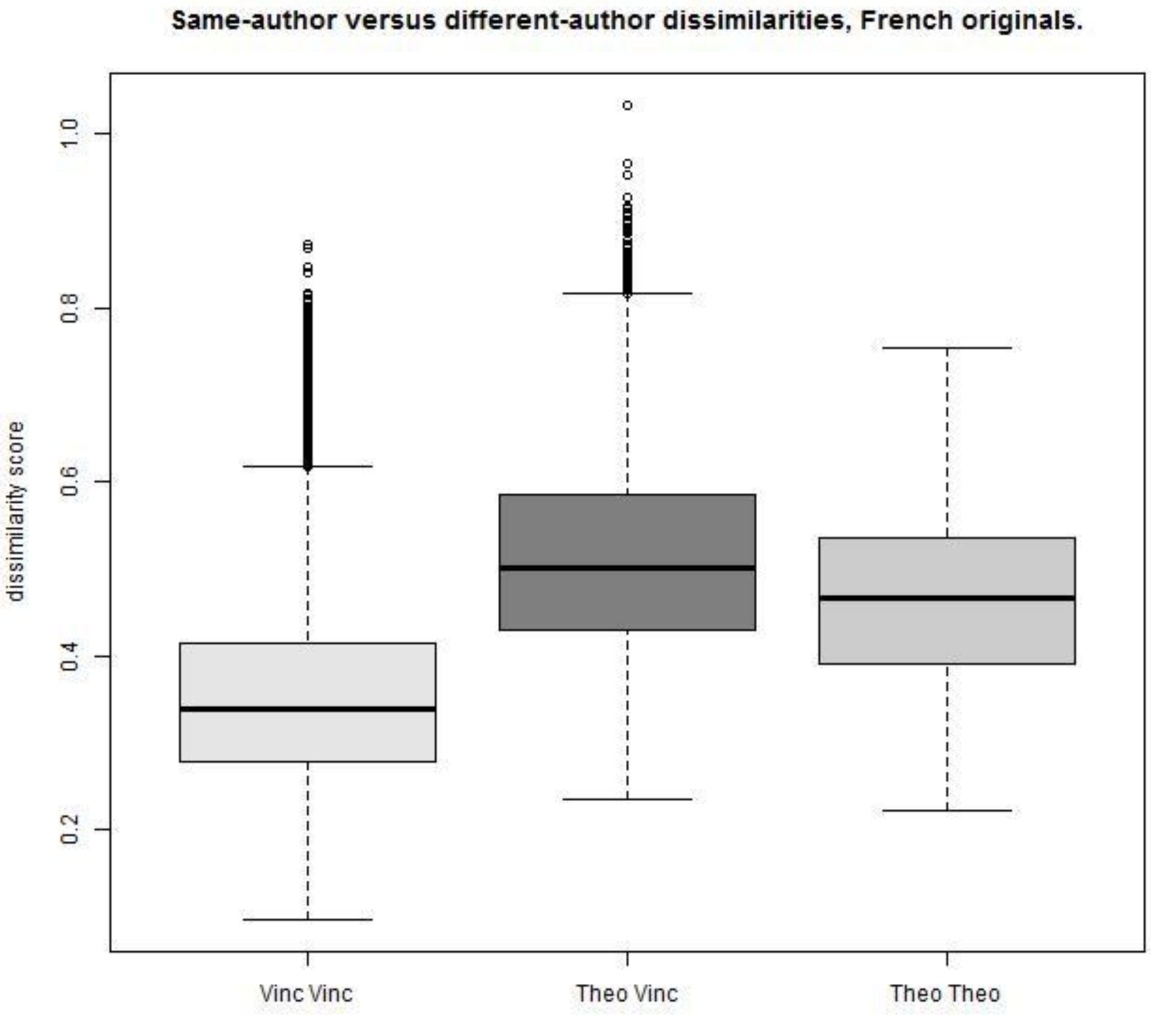

Figure 1. Inter-text dissimilarities within authors and between authors, French originals.

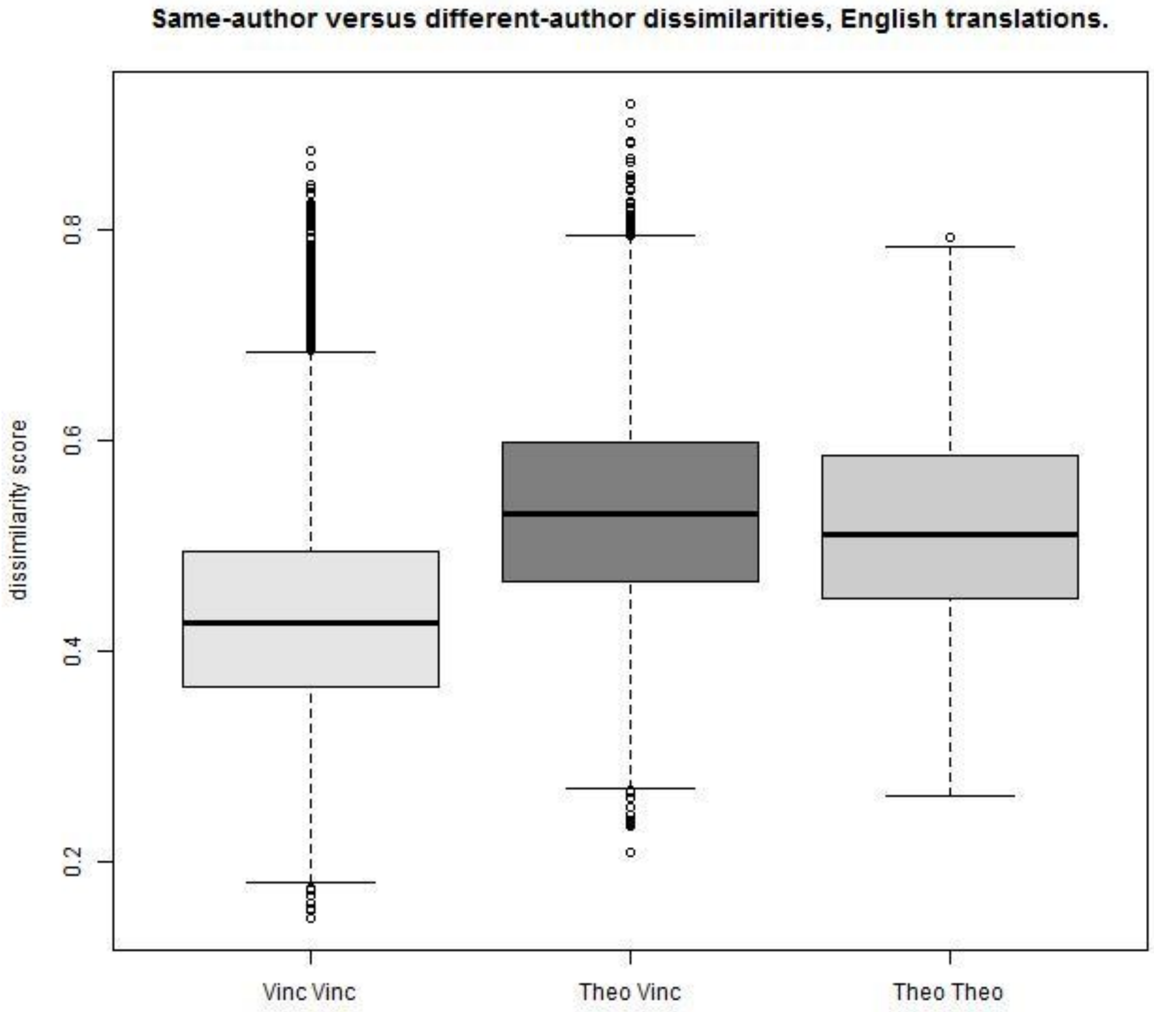

Figure 2. Inter-text dissimilarities within authors and between authors, VGM English translations.

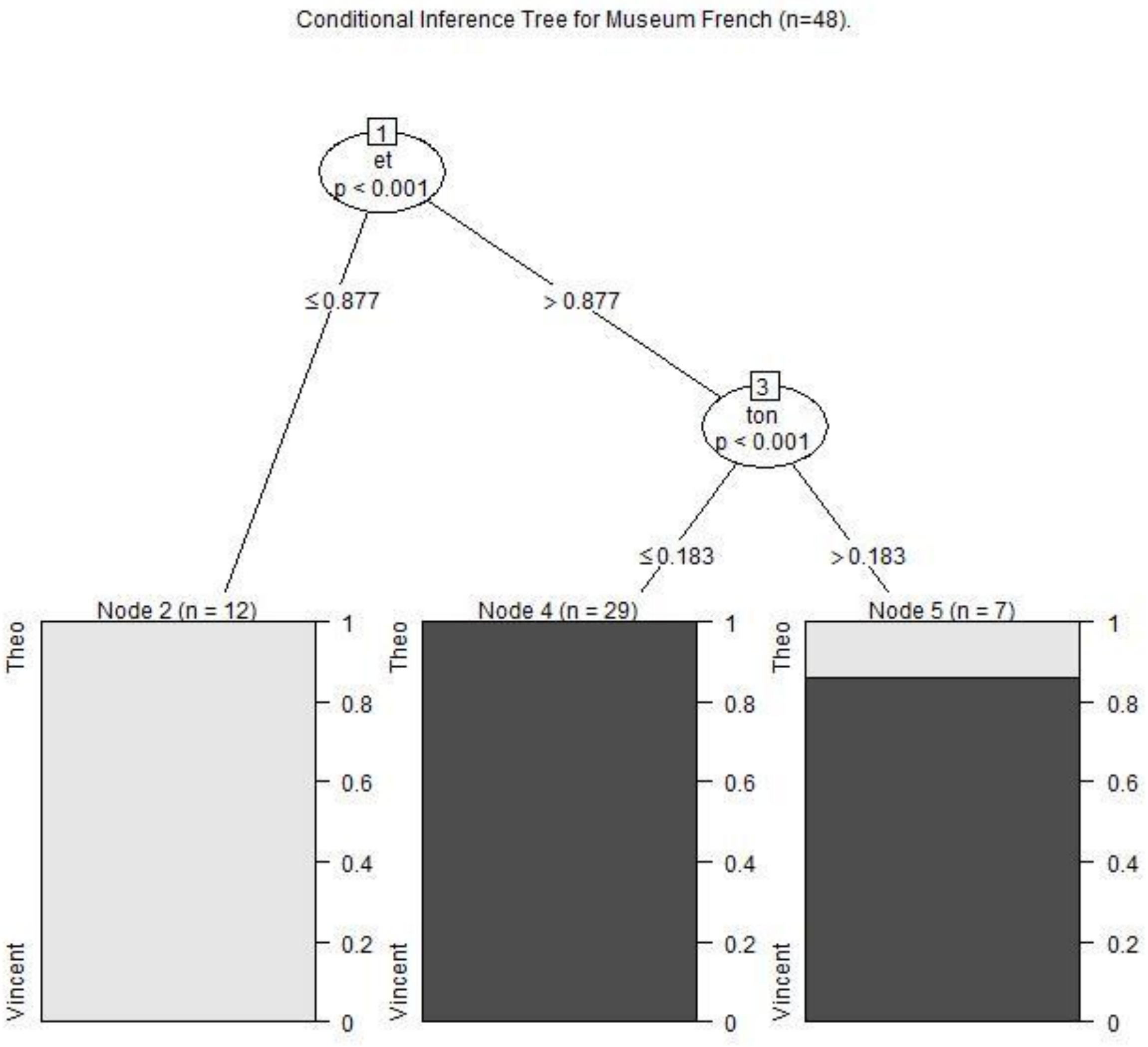

Figure 3. Discrimination tree for Theo versus Vincent, French originals.

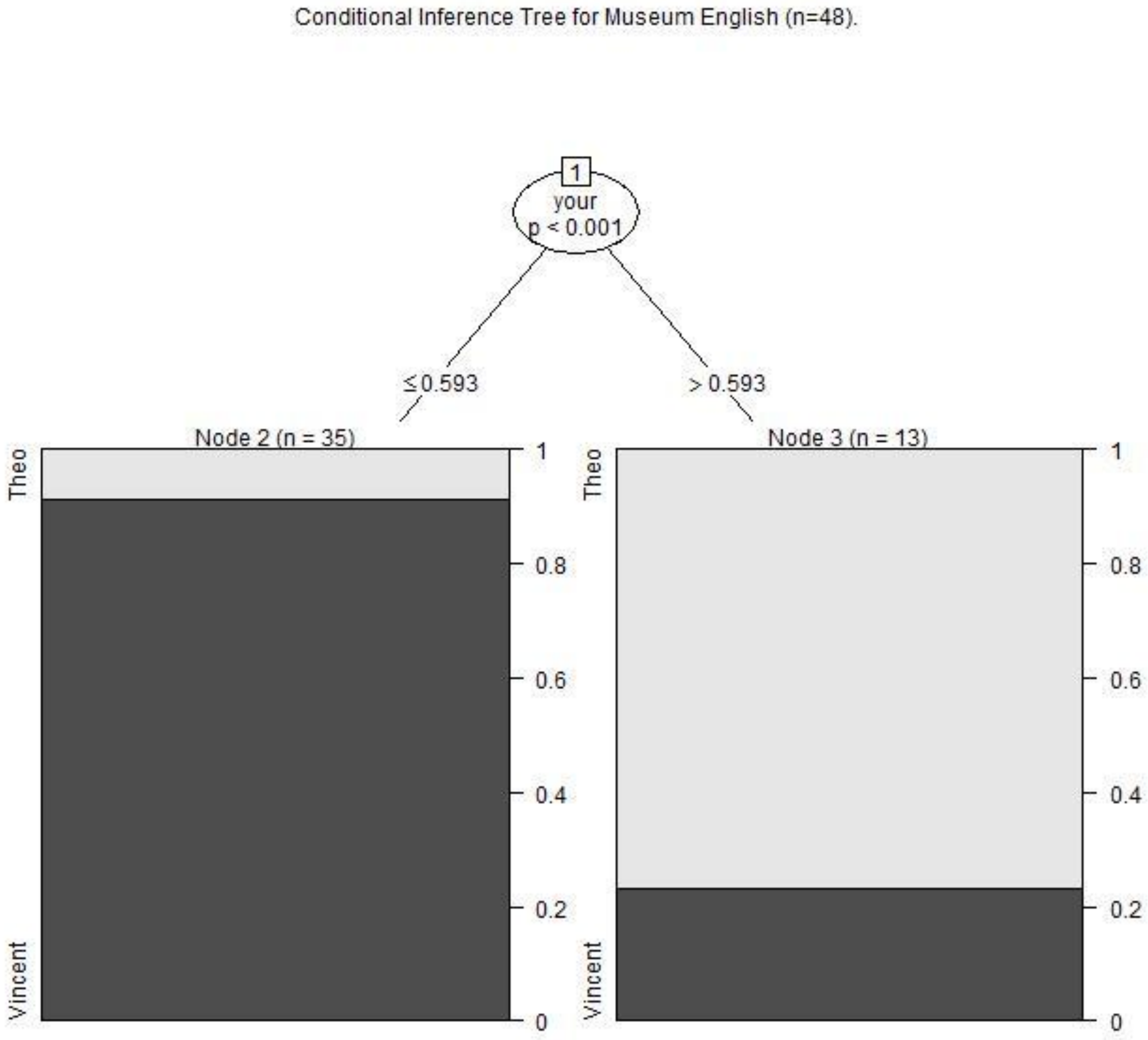

Figure 4. Discrimination tree for Theo versus Vincent, VGM English translations.

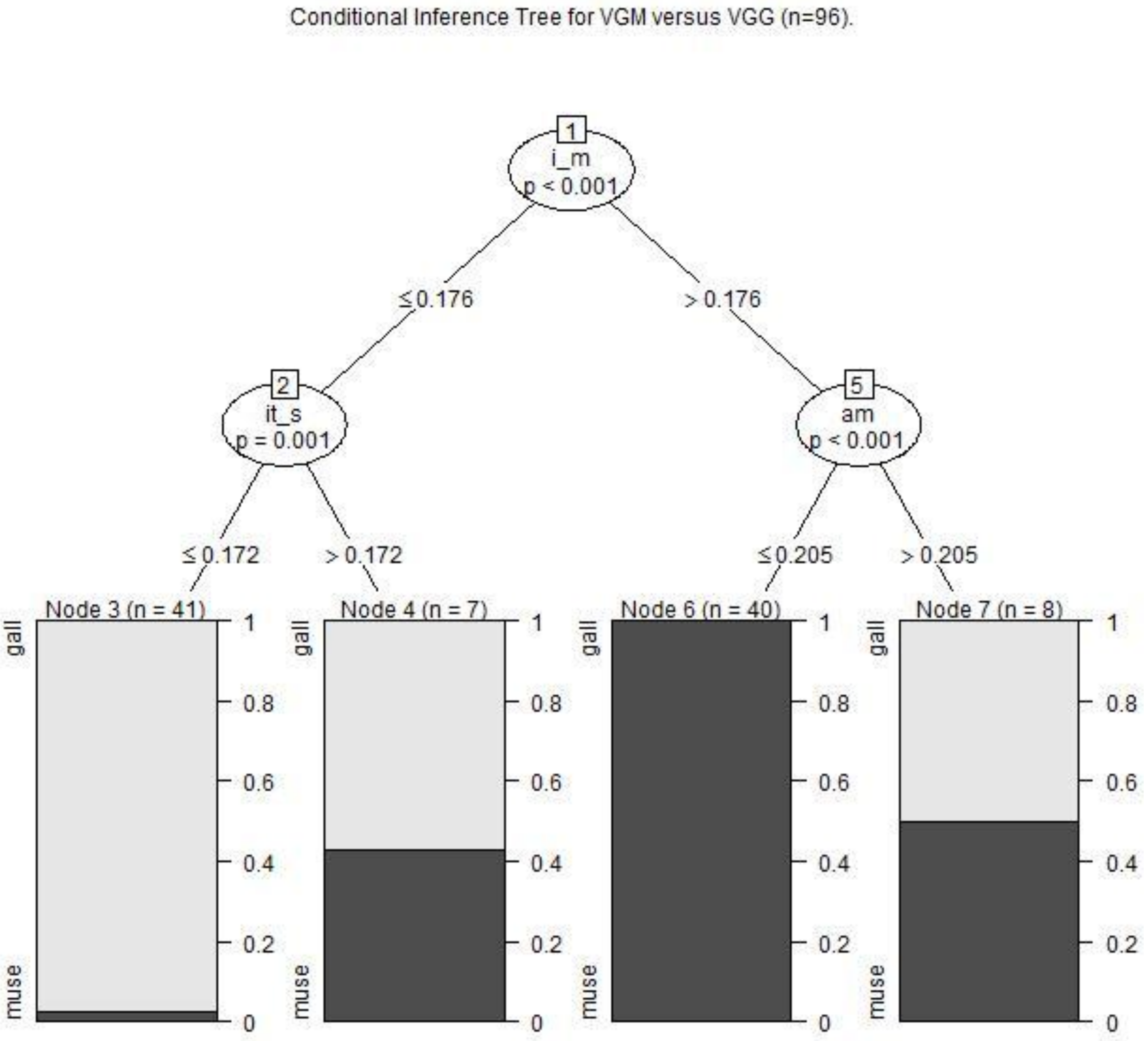

Figure 5. Discrimination tree for Museum versus Gallery translations.