A clustering approach for translationese identification

Sergiu Nisioi
Faculty of Mathematics and
Computer Science,
University of Bucharest
sergiu.nisioi@gmail.com

Liviu P. Dinu
Centre for Computational
Linguistics, Bucharest
ldinu@fmi.unibuc.ro

Abstract

Our paper is concerned with investigating the impact of translationese on the novels of a bilingual writer and asking whether one could determine the authorship of a translated document. The main part of our paper will be centered on selecting a good set of lexical features that can be considered characteristic for an author. We used in our research the novels of Vladimir Nabokov, a bilingual author, who wrote his works in both Russian and English. Each text is represented by a vector of function words. We are interested in determining how the results vary across different feature sets and which feature set could be considered the most representative. In order to inspect our results we used a hierarchical clustering method and draw conclusions based on the most frequent result.

1 Introduction

The term "translationese" proposed by Gellerstam (1986) currently means the entire sum of linguistic characteristics (Hansen, 2003) that a translation exhibits in comparison to a text written natively in a language. The existence of translationese has been discussed and more recently various methods (Koppel et al., 2011; Ilisei et al., 2010) for identifying translationese have been devised.

In the same context, an interesting discussion regards the equivalence in style between the translated and original text. As Boase-Beier (2006) suggests, among other factors, the stylistics of a translation is highly related to the choices made by the translator in re-creating the original style, the translator having a specific "fingerprint" (Wang and Li, 2012). Our concern is investigating the impact of translationese on a bilingual writer and asking whether one could determine the authorship of a translated document. The problem of authorship attribution is postulated on the grounds that the human stylome exists. The stylome is defined as "a linguistic fingerprint that can be measured, is largely unconscious, and is constant" (van Halteren et al., 2005). A fairly large amount of literature is dedicated to authorship problems and extensive overviews are provided by Juola (2006) or Stamatatos (2009).

We are mostly interested in finding the lexical features that can be used to discriminate or to characterize original and translated documents and once these words are presumably found, what is their role in authorship attribution for such documents? The main part of our paper will be centered on selecting a *good* set of lexical features to detect translations in a corpus of original documents.

In order to investigate our problems we have constructed two corpora from the novels of Vladimir Nabokov: a Russian corpus containing original Russian works and translations from English and a second corpus containing English original works and translations from Russian. The details with respect to each work included are to be found in Section 2. The fact that Vladimir Nabokov was bilingual (McKenna et al., 1999) certainly affects the interpretation of the results. On one side there exists a difference of style between author and translator and secondly a translation preserves enough translationese to make it different from any other original text.

For lexical feature sets, two quality criteria are commonly used in literature: one, the lexical features should have a relatively high frequency. Rybicki and Eder (2011) have reported better results with high frequency words. The second criterion is to consider function words instead of content words. Function words do not contain information about the topic of the text and are used un-

Proceedings of Recent Advances in Natural Language Processing, pages 532–538,
Hissar, Bulgaria, 7-13 September 2013.

consciously revealing important psychological aspects (Chung and Pennebaker, 2007). Moreover, these words are used to tie phrases and help making stylistic constructs that can be specific to one author. These two criteria were first attested by Mosteller and Wallace (1963) and remained an important decision factor until today.

2 Corpus

We have focused our analysis on the novels of Vladimir Nabokov by using two main corpora: one in Russian containing the original Russian novels (written before 1940) together with the translations of Nabokov's original English novels, and a second one containing the original English novels (written after 1940) together with various translations of his novels into English. Except for *Lolita* all the translations into Russian are done by Sergey Ilyin. This does not influence our results greatly for two main reasons: one *Lolita* is never clustered among the Russian novels although it was translated by Nabokov and two *Dar* is not always clustered among the Russian novels although it was originally written in Russian.

Traces of the author should exist in all the English translations since V. Nabokov collaborated in translating them.

Finally, the size of our Russian corpus reached 1,062,594 words and the size of the English corpus a smaller 904,712 number of words. Our hypothesis is that the original novels of Nabokov will be clustered separately from the translated ones without regarding the language.

We are confident that the works of Nabokov constitute two significant corpora on two different languages that are meaningful for comparisons.

In order to answer our second problem regarding the attribution of a translation we have added additional writers to our experiment. These authors are Alexey Tolstoy, Lev Tolstoy, Fyodor Dostoyevsky, Iury Olesha, Valery Bryusov, Ilf and Petrov, Boris Pasternak, Andrey Bely and Ivan Turgenev.

3 Using ranks and classification

Since we have a relatively small number of documents of significant size each (in both the Russian and the English corpus), we believe that hierarchical clustering will offer sufficient details to pursue our investigation and that it could determine homogenous groups providing additional information in comparison with a simple binary classification task.

We have used Burrows' Δ to calculate a similarity matrix as input for the clustering algorithm. This measure enjoyed a lot of attention (Argamon, 2008), producing results comparable with the ones of learning methods on authorship attribution. In our case, the use of Δ will be to distinguish between translated texts and original ones.

The equation of Δ is:

$$\Delta^{(n)}(D, D') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i}|f_i(D) - f_i(D')| \quad (1)$$

Where n is the size of our vectors or the number of words from our feature set, D and D' two vector documents, σ_i the standard deviation of word i in the whole corpus, f_i the frequency of word i in D and D'.

We can easily observe that if we remove the constant fraction $1/n$, the value of Δ is actually equal with the $l1$ distance between z-scores, defined as $\sum_{i=1}^{n}|z(x_i) - z(y_i)|$ where z is the z-score of a word equal to $z(x_i) = \frac{f_i - \mu_i}{\sigma_i}$.

In order to visualize the results we have used an $l1$ norm (Dinu and Nisioi, 2012) modified version of the hierarchical clustering algorithm proposed by Szekely and Rizzo (2005). Their algorithm is a bottom-up approach to generalize Ward's minimum variance method (Ward, 1963) by defining a cluster distance and objective function in terms of Euclidean distance. In addition it has the ability of identifying clusters with nearly equal centers and it was successfully used for classifying diseases (Szekely and Rizzo, 2005).

Dinu and Popescu (2008) introduced a ranking operation on the frequency vectors of each documents with the purpose of eliminating outliers (produced by large vs. small frequencies of words) thus making the distances between texts measurable and more stable. As a result, it produced confident results in a case of pastiche detection on Romanian (Dinu et al., 2012). We will further test this approach by applying it on Russian on a bilingual author in a different situation.

A ranking of a vector of n words is a mapping $\tau : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., n\}$ where $\tau(f(i))$ will represent the place (rank) of the frequency (as in Equation 1) of the word indexed as i. If $\tau(f(i)) < \tau(f(j))$ then the word i is more frequent than word j. In our case of using Δ, no

Russian	**Number of tokens**	**English**	**Number of tokens**
Mashenka (1926) (O)	26,173	*Mary* (1970) (T: Michael Glenny and V. Nabokov)	34,906
Korol' Dama Valet (1928) (O)	57,123	-	-
Zashchita Luzhina (1930) (O)	54,013	*The (Luzhin) Defence* (1964) (T: Michael Glenny and V. Nabokov)	75,417
Podvig (1932) (O)	54,372	-	-
Camera Obskura (1933) (O)	45,245	*Laughter in the Dark* (1938) (T: V. Nabokov)	62,006
Otchayanie (1934) (O)	47,199	-	-
Priglasheniye na kazn (1936) (O)	42,429	*Invitation to a Beheading* (1959) (T: D. Nabokov and V. Nabokov)	60,195
Dar (1938) (O)	116,330	-	-
Podlinnaya zhizn Sebastyana Nayta (T: S. Ilyin)	54,180	*The Real Life of Sebastian Knight* (1941) (O)	62,390
Pod znakom nezakonnorozhdënnykh (T: S. Ilyin)	60,035	*Bend Sinister* (1947) (O)	73,075
Lolita (T: V. Nabokov)	117,287	*Lolita* (1955) (O)	117,185
Pnin (T: S. Ilyin)	48,984	*Pnin* (1957) (O)	52,628
Blednoye plamya (T: S. Ilyin)	81,816	*Pale Fire* (1962) (O)	85,164
Ada (T: S. Ilyin)	168,103	*Ada or Ardor: A Family Chronicle* (1969) (O)	181,346
Prozrachnyye veshchi (T: S. Ilyin)	25,898	*Transparent Things* (1972) (O)	29,073
Smotri na arlekinov! (T: S. Ilyin)	63,407	*Look at the Harlequins!* (1974) (O)	71,327
Russian Total	1,062,594	***English Total***	904,712

Table 1: Left, we have the Russian novels in original(O) and the translations of Sergey Ilyin. The size in words of the Russian corpus is 1,062,594. Right, we have the English novels in original together with a subset of translations from Russian. We could not obtain all the equivalent translations, the Eglish corpus having a smaller size of 904,712 words.

difference is made if the ordering relation is increasing or decreasing (Dinu and Nisioi, 2012).

This is our last operation onto the matrix of similarities before inputting it in the clustering algorithm. We have linearized the matrix (converted it into a vector of measurements obtained in this case from computing delta between each pair of novels). Each value was replaced by its tied rank in the entire vector (Dinu and Popescu, 2008). Then we have reordered the values back into the initial matrix. The reason for this operation is that small distances increase between each other and large distances decrease making the method more robust.

4 Feature set

On the English corpus, we have tested the feature set proposed by Mosteller and Wallace (1963) consisting from the words: a, been, had, its, one, that, was, all, but, has, may, only, the, were, also, by, have, more, or, their, what, an, can, her, must, our, then, when, and, do, his, my, shall, there, which, any, down, if, no, should, things, who, are, even, in, not, so, this, will, as, every, into, now, some, to, with, at, for, is, of, such, up, would, be, from, it, on, than, upon, your.

In this case each document becomes a vector of size 70 in which each entry represents the frequency for the corresponding feature. The text was preprocessed to remove punctuation marks and other signs.

4.1 Feature selection

The majority of the studies rely on the principle of the most frequent words from the corpus. However, finding the exact number of the most frequent words is subject of extensive debate which dates since the study of Mosteller and Wallace (1963). Rybicki and Eder (2011) correlate the number with the language properties, other studies (Hoover, 2004; Smith and Aldridge, 2011) eliminate certain classes of words and Jockers and Witten (2012) researched optimal thresholds for word frequencies. This problem persists in every case of word usage (Koppel et al., 2007) method. Overall, the problem of selecting an objective feature set does not have a straight forward solution.

For Russian, we have introduced a process of selecting a feature set based on quantitative aspects of the results produced.

We start with the premise that the clustering results are representative with respect to the distances measured. This is assured by the $l1$ change introduced by Dinu and Nisioi (2012) and by replacing the values with ranks inside the similarity matrix. Our comparisons depend on the clustering results obtained. Using various different lists of the most frequent function words, we have executed a computational process to produce for each list a dendrogram.

The outline of the algorithm is presented below:

Algorithm 1 - *for selecting the best feature set based on measured quality*

```
1.  let F = the function words
from a Language
```

```
2. sort F decreasing by frequency in the corpus
3. exclude from F the words that are missing in at least one document from the corpus
4. let $n$ be the size of F and $h = n/2$
for all i from $h$ to $n$ do
    4.1 $F_i$ = the first $i$ elements from $F$ {the same as the first $i$ function words from the corpus}
    4.2 for each document in the corpus construct vectors of frequencies using the list $F_i$
    4.3 for each vector representation of a document replace frequencies by ranks {as detailed in Section 3}
    4.4 let M = matrix obtained from $\Delta$ computed between each vector pair
    4.5 linearise M, replace the values by ranks (similar with 4.3), reconstruct M as a matrix
    4.6 let $R_i$ = dendrogram obtained using hierarchical clustering algorithm with input M
    4.7 let $U_i$ = un-weighted tree of dendrogram $R_i$
    4.8 let counter[$U_i$] = counter[$U_i$] + 1 {increase the cardinal of each equivalence class generated by $R_i$}
    4.9 record that the feature set $F_i$ produced the result $R_i$
end for
5. let RESULT = $R_i$ for which counter[$U_i$] is maximum
6. let FEATURE_SET = minimum set $F_i$ which generated $R_i$
```

The first step is to retrieve all the Russian conjunctions, determiners, particles, prepositions, pronouns and adverbs (function words) from ru.wiktionary.org. We have experimented with different classes of function words from Russian and the best results were obtained by partially removing pronouns. Previous studies like the one of Hoover (2004) also suggest this operation. The second step is to order the list descending, by frequency of appearance in the entire corpus. The third step is to select from this list only the words that appear individually at least once in each document.

The fourth step is about selecting a lower limit from which to start comparing the first, most frequent function words. Starting from the half of the entire function words list (notated as F) can be a good decision, especially if the list has a significant large size. It is not entirely clear what is the minimum number of words needed to characterize the style of a text. This is why the starting value of h will be left for the user to decide according to the case. In our case, let $F_{n/2}$ be the first $n/2$ function words sorted descending by frequency in corpus. Then for each $n/2 < i \leq n$ we create a set F_i by adding consecutively one more word found at position i in the entire set F. Thus the comparisons will be made between results computed with the lists of the first i most frequent function words F_i.

Moreover, for Russian the function words have a relatively high number of declensions, so in order to correctly count the features, all the text was POS-tagged and lemmatized using TreeTagger (Schmid, 1995). On English, this operation was not necessary since the list of words provided by Mosteller and Wallace contains un-lemmatized words (e. g. “been”, “had”, “were”, etc.).

The hierarchical clustering algorithm has as input a similarity matrix and the result is illustrated by a binary tree called dendrogram, as we can observe in Figure 1.

If we ignore the distances that the edges (also called weighted links) have between clusters, we obtain a simple binary tree that illustrates only the arrangements of the clusters. We will consider two dendrograms to be equivalent if their un-weighted binary trees are identical. This means that two dendrograms are equivalent if the arrangements of the clusters are identical. Other equivalence relations can be defined at this point, depending on the size of the corpus and the works which need to be emphasized.

Roughly, the algorithm constructs for each feature set the vector representation of the documents, replaces frequencies with ranks then computes the similarity matrix using Δ, cluster the documents,

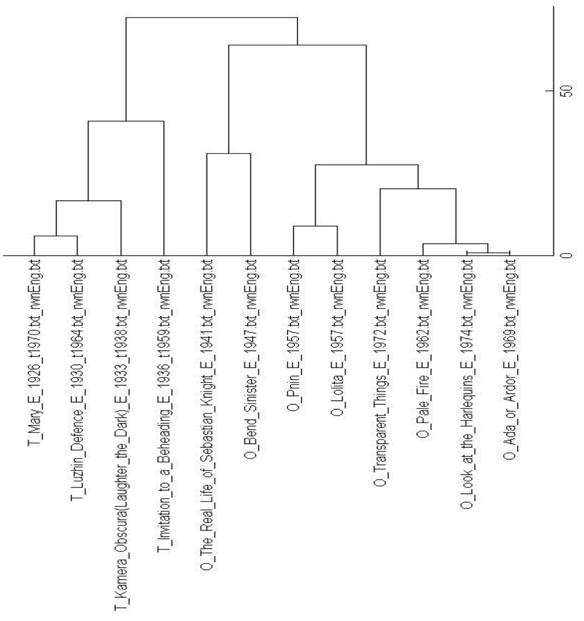

Figure 1: English corpus clustering result with the function words of Mosteller and Wallace. Translations (prefixed with T) are clustered separately from all the original works (prefixed with O). Moreover, we can observe a very small difference between the image with the original English novels and S. Ilyin's translations into Russian in Figure 2.

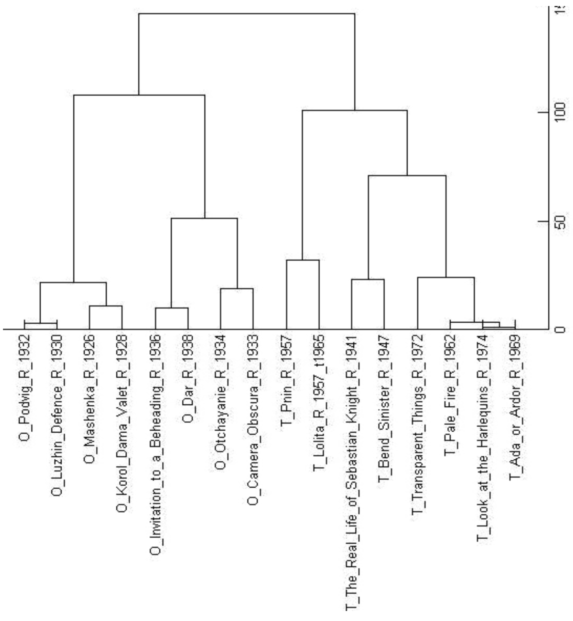

Figure 2: The most frequent Russian corpus clustering result. Translations (prefixed with T) are clustered separately from all the original works (prefixed with O). Furthermore, the original Russian novels are clustered chronologically two by two: 1932 - 1930, 1926 - 1928, 1936 - 1938, 1934 - 1933. The translations are clustered in a similar way as the originals in Figure 1.

obtains the dendrogram and based on the relation defined early it groups the equivalent results. Finally, we obtain, based on the arrangements of the clusters, different classes of results (equivalence classes).

A result is "better" (has a greater degree of quality) than another if the equivalence class of the result is larger than the equivalence class of the other.

The best result is the one with the most often produced un-weighted tree for various feature sets. Since the algorithm produces the same un-weighted tree with more words, we could just eliminate the surplus and keep only the smallest number of words.

Thus, from all the feature sets that produced the best result, we consider the smallest feature set to be the most representative for one specific corpus of text.

This criterion expresses the general tendency of the documents to be clustered in a particular way under an entire class of feature sets.

5 Results

In Figure 1 and Figure 2 we can observe that there are two main clusters of original (prefixed with O) and translated documents (prefixed with T). For Russian, the most frequent result was found with the minimum size of the list being $n = 94$ words. On English it produced the same RESULT as with the Mosteller and Wallace features Figure 1.

Using frequencies instead of ranks on both Russian and English failed to validate the hypothesis regarding translationese detection.

The RESULT of the algorithm can be seen in Figure 2 and the final FEATURE_LIST of 94 words, computed for Russian is: и, в, не, что, на, быть, с, как, а, это, но, к, по, же, так, то, из, за, у, бы, весь, от, о, только, да, уже, вот, когда, даже, до, или, для, если, другой, вдруг, время, ни, ли, чтобы, раз, во, под, со, чем, кто, два, без, потому, при, тогда, между, надо, через, над, сейчас, можно, будто, об, больше, всегда, хотя, перед, про, всякий, случай, именно, хоть, много, точно, довольно, пока, куда, давно, иногда, ко, иной, быстро, долго, едва, мало, завтра, также, сквозь,

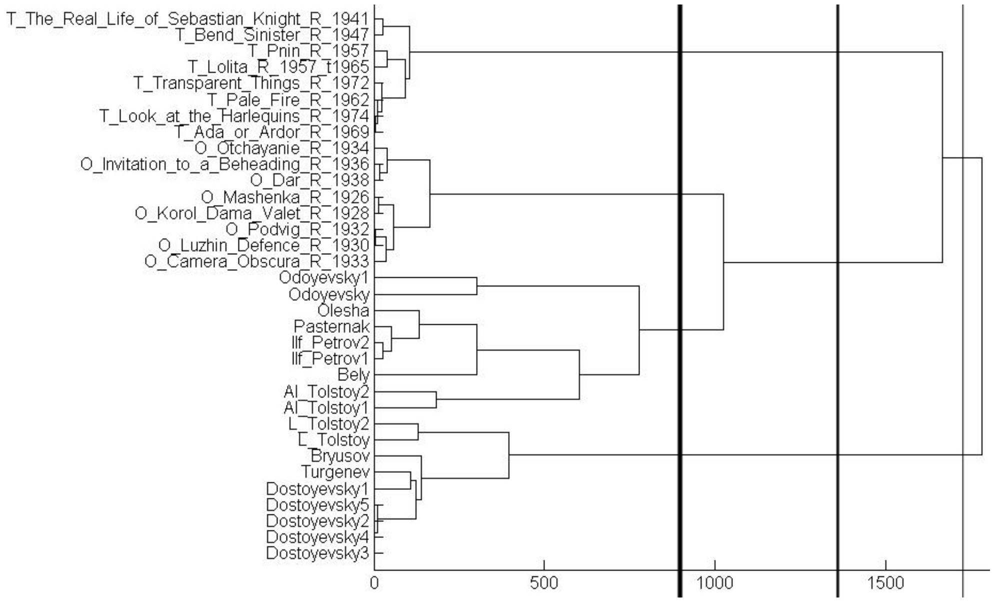

Figure 3: The result obtained from the first 94 function words (computed by the algorithm described) from the entire corpus. Nabokov is separated from other Russian authors and the translations are also separated. For each author the corresponding works are grouped in the same cluster.

мимо, домой, против, напротив, около, далеко, видно, вокруг, гораздо, вон, весело

Observation: Sometimes the counter of the "the best" result and the counter of the "second best" result could have similar values but this was not the case for us. In those situations, we would advise to analyze the differences between the results. Moreover, in the case when a large number of files are analyzed the counters could be small and the clusters could differ by a small shift between one and other. We indicate choosing a different equivalence relation in this scenario.

Using the same feature set deduced early, we have obtained just the clustering result on Russian. Figure 3 is relevant in this sense. A first cut of this dendrogram (the rightmost thin vertical line) indicate that the early Russian novelists included (Dostoyevsky, Bryusov, L. Tolstoy, Turgenev) are clustered separately from all the other authors. A second cut (the middle-sized vertical line) indicate an answer to our second problem - if we can attribute a translation to an author. Giving the inter-cluster distance, we find the original works of Nabokov (prefixed with O) as being closest to the cluster containing translated works (prefixed with T). The third cut (the leftmost thick vertical line) suggests that translations (prefixed with T) are clustered separately from all the other Russian novels. This fact enforces the theory under which translations have distinctive features from text written natively in a certain language. Nevertheless, all the texts of various writers (including Nabokov) are clustered together, confirming the possibility of attributing translated documents to an author.

6 Conclusions and future work

We have reconsidered from a quantitative perspective the works of a bilingual writer. We have created a list of function words for Russian with respect to some quality factors described (frequency of words, frequency of results, word types and rankings) and tested it in a larger context by extending the initial corpus of Nabokov's works, see Figure 3. In the same figure the works of the other authors are grouped accordingly. The cluster of original works of Nabokov is the closest cluster to the translated from English documents (with respect to the inter-cluster distance). It seems that there is a tight relation between translationese identification and authorship attribution since features normally used to characterize the style of an author can be used to classify translated versus original documents.

As for of attributing a translation we can confirm that it is possible in a certain degree of fuzziness, Figure 3. We have to also consider fact that Nabokov, at the time of writing in English had assimilated the language perfectly as suggested by Gorski (2010).

The immediate priority is enlarging the English corpus for testing and extending the methods presented. Another chapter of interest is related to finding the linguistic resorts behind these feature sets and what other properties do they present with respect to the corpus. Analyzing the similarities between different translations of the same work is also on top of our list.

References

Argamon, S. 2008 Interpreting Burrows' Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2) 131-147

Boase-Beier, J. 2006 *Stylistic Approaches to Translation* Manchester: St Jerome, (2006).

Burrows, J.F. 2002 *Delta: A measure of stylistic difference and a guide to likely authorship.* Literary and Linguistic Computing 17 267–287

Chung, K.C. and Pennebaker, J.W. 2007 *The psychological functions of function words* Psychology Press, New York. 343–359

Dinu, L.P., Niculae, V., Şulea, O.M. 2012 Pastiche detection based on stopword rankings: exposing impersonators of a romanian writer. In: *Proceedings of the Workshop on Computational Approaches to Deception Detection. EACL 2012* Stroudsburg, PA, USA, Association for Computational Linguistics, 72–77

Dinu, L.P. and Nisioi, S. 2012 Authorial Studies using Ranked Lexical Features *Demos Proceedings of COLING 2012*, Mumbay, 125–130

Dinu, L.P. and Popescu, M. 2008 Rank distance as a stylistic similarity *Proceedings of COLING 2008*, Manchester, ELRA, 91–94

Hoover, D.L. 2004 Testing burrows's delta *Literary and Linguistic Computing*, 19, 453–475

Gorski, B. 2010 Nabokov vs. Набоков: A literary investigation of linguistic relativity *VESTNIK, THE JOURNAL OF RUSSIAN AND ASIAN STUDIES*

Gellerstam, M. 1986 Translationese in Swedish novels translated from English *Translation Studies in Scandinavia*, Lund: CWK Gleerup

Hansen, S. 2003 *The Nature of Translated Text* Saarbrucken: Saarland University

van Halteren, Hans and Baayen, R.H and Tweedie, F. and Haverkort, M. and Neijt, A. 2005 New machine learning methods demonstrate the existence of a human stylome *Journal of Quantitative Linguistics*, 65–77

Ilisei, I. and Inkpen, D. and Corpas Pastor, G. and Mitkov, R. 2010 Identification of translationese: a machine learning approach *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing*, 503–511

Jockers, M.L. and Witten, D.M. 2012 A comparative study of machine learning methods for authorship attribution *Literary and Linguist Computing*, 215–223.

Juola, P. 2006 Authorship Attribution *Foundations and Trends in Information Retrieval*, Vol. 1, Nr. 3, 233-334

Koppel, M., Schler, J., Bonchek-Dokow, E. 2007 Measuring differentiability: Unmasking pseudonymous authors *Journal of Machine Learning* Res. 8, 1261–1276

Koppel, M., Ordan., N. 2011 Translationese and its dialects *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* Volume 1, 9, 1318–1326

McKenna, W., Burrows, J., Antonia, A. 1999 Beckett's "molloy": Computational stylistics and the meaning of translation *Variété: Perspectives in French Literature*, Society and Culture. Studies in Honour of Kenneth Raymond Dutton 79–92

Mosteller, F. and Wallace, L.D. 1963 Inference in an authorship problem *Journal of the American Statistical Association*, 58 275–309

Rybicki, J. and Eder, M. 2011 Deeper Delta across genres and languages: do we really need the most frequent words? *Journal of Litearary and Linguistic Computing*, 26, 3, 315 – 321

Schmid, H. 1995 Improvements in Part-of-Speech Tagging with an Application to German *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland

Smith, P.W.H. and Aldridge, W. 2011 Improving authorship attribution: Optimizing burrows' delta method*. *Journal of Quantitative Linguistics 18*, 63–88

Stamatatos, E. 2009 A survey of modern authorship attribution methods *J. Am. Soc. Inf. Sci. Technol.*, Vol. 60, Nr. 3, 538–556

Szekely, G.J. and Rizzo, M.L. 2005 Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 151 – 183

Wang, Q. and Li, D. 2012 Looking for translator's fingerprints: a corpus-based study on chinese translations of ulysses. *Literary and Linguistic Computing*, 27, 81–93

s Ward, J.H. 1963 Hierarchical grouping to optimize an objective function. *J. of American Statistical Association*, 301, 236–244