Unsupervised Classification of Translated Texts

Sergiu Nisioi(✉)

Center for Computational Linguistics, University of Bucharest, Bucharest, Romania
sergiu.nisioi@gmail.com

Abstract. In our paper we investigate the possibility to use an unsupervised classifier to automatically distinguish between the translated and original novels of a multilingual writer (Vladimir Nabokov) and to determine whether the authorship of a translated document can be achieved. We employ a rank-based document vector representation using only function words as features. To extract the results, we propose a generalization of Ward's hierarchical clustering method that is compatible with any similarity metric.

1 Introduction

The research of automatic methods to measure stylistic similarities between texts has a long history, but one of the first successful studies in this direction is that of Mosteller and Wallace [22]. Their approach combined statistical models with linguistic information to infer the authorship of disputed Federalist papers. The linguistic information comprised of certain word categories extracted from the documents, concluding that the class of function words can act as an author's *fingerprint* for a text.

In our work we are interested to observe whether the style of an author can be preserved by translation, given that the style is defined by an author's use of function words. In this sense we compare the translated texts of a multilingual author (T) with the works originally written by the same author (O).

During translation, grammatical structures of the source language (O) can get printed unintentionally to the target/translation (T). In translation theory, this phenomenon is termed interference. *Language transfer* is a similar phenomenon in language acquisition theory which describes the influence carried from the mother tongue to the utterances in other languages spoken by an individual. Since we are investigating the works of a multilingual writer, both of these phenomena are likely to appear in the texts. Previous machine classification studies investigating interference [14,27,33] or language transfer [26,32] indicate that function words can be reliable, topic independent features that evidence these phenomena.

The author in our discussion is Vladimir Nabokov, a multilingual Russian-American novelist who wrote most of his Russian novels living in exile in Europe and switched to English after his departure in USA. We have constructed two significant Russian-English corpora containing both the original and the translated novels on which we attempt to apply a generalization of Ward's clustering method.

C. Biemann et al. (Eds.): NLDB 2015, LNCS 9103, pp. 323–334, 2015.
DOI: 10.1007/978-3-319-19581-0_29

2 The Nabokov Corpus - Interference and Language Transfer

The corpus is compiled out of ten Russian (O) novels and eight English (O) novels together with the corresponding translations (T) of each. The English translations have a better chance to preserve the original *fingerprint* of the author since he supervised and contributed to almost every work, while the Russian translations are more homogeneous, being translated by Sergey Ilyin. If the author is "more present" in the English translations, then we should expect the classifications to contain a larger degree of confusion between T and O in the English corpus. On the Russian side, *Lolita* is the only work translated into Russian by the author. Table 1 contains the details with respect to each novel included.

Both interference and language transfer could be present in Nabokov's translations. It is difficult to asses the amount of language transfer for a trilingual (Russian, English, French) author whose first reading language was probably [24] English. On one hand, Gorski [10] analyzing Nabokov's autobiographical works concludes that our author had near-native skills in English. On the other hand, from a second language acquisition perspective, Selinker and Rutherford [29] claim that a so-called fossilization intervenes for language learners. *Fossilization* designates the permanent cessation of target language (TL) learning before the learner has attained the TL norms at all levels of linguistic structure.

If such would be the case, then any of Nabokov's English novels as well as his translations into English would be, in fact, utterances of a fossilized interlanguage [29] - an independent linguistic system different from the mother tongue of an individual and from the languages acquired. Given the series of audio recordings of his English interviews, we can trace the presence of the open-mid front rounded vowel and other French specific phonological patterns [11] in a mix of British and Russian pronunciation of the voiced alveolar trill [r]. In this sense, we can observe an obvious effect of fossilization of the interlanguage at the phonological level. Nabokov himself claimed at the end of the English version of *Lolita* that he abandoned *my natural idiom, my untrammelled, rich, and infinitely docile Russian tongue for a second-rate brand of English* [23].

We are inclined to believe the translations in our corpus are *literal*, as the author puts it: *rendering, as closely as the associative and syntactical capacities of another language allow, the exact contextual meaning of the original. Only this is true translation* [25]. Under this assumption, interference should be visible in every translation that he approved or collaborated in English or Russian.

Although the works are written many years apart, there is no literary hypothesis to suggest that Nabokov went through a change of style after starting to write in English. Furthermore, the corpus is semi-aligned and the translators are varied, if similar results are extracted from English and Russian, we can be confident that the differences emerge due to a clear distinction between translator and author, including a possible connection with the language transfer phenomenon.

Table 1. The Russian-English corpora are represented in this table, on the left column the titles of original (O) and translations (T) are provided in Russian. The right column contains the English title and the translators who collaborated for that work. The year of writing/translating a certain novel is marked between parentheses. The size is measured as the number of tokens for each work.

Russian	Size	English	Size
Mashenka (1926) (O)	25,131	*Mary* (1970) (T: Michael Glenny and V. Nabokov)	34,359
Korol' Dama Valet (1928) (O)	55,149	*King, Queen, Knave* (1968) (T: Dmitri Nabokov)	83,975
Zashchita Luzhina (1930) (O)	52,173	*The (Luzhin) Defence* (1964) (T: Michael Glenny and V. Nabokov)	75,417
Sogliadatai (1930) (O)	16,007	*The Eye* (1965) (T: Dmitri Nabokov)	22,715
Podvig (1932) (O)	54,372	*Glory* (1971) (T: Dmitri Nabokov)	67,314
Camera Obskura (1933) (O)	43,566	*Laughter in the Dark* (1938) (T: V. Nabokov)	56,937
Otchayanie (1934) (O)	42,811	*Despair* (1965) (T: Vladimir Nabokov)	65,412
Priglasheniye na kazn (1936) (O)	40,434	*Invitation to a Beheading* (1959) (T: D. Nabokov and V. Nabokov)	56,081
Dar (1938) (O)	105,528	*The Gift* (1963) (T: Dmitri Nabokov)	115,265
Volshebnik (1939) (O)	12,106	*The Enchanter* (1986) (T: Dmitri Nabokov)	25,821
Podlinnaya zhizn Sebastyana Nayta (T: S. Ilyin)	49,435	*The Real Life of Sebastian Knight* (1941) (O)	62,390
Pod znakom nezakonnorozhdënnykh (T: S. Ilyin)	56,959	*Bend Sinister* (1947) (O)	73,075
Lolita (T: V. Nabokov)	107,271	*Lolita* (1955) (O)	117,185
Pnin (T: S. Ilyin)	46,584	*Pnin* (1957) (O)	52,628
Blednoye plamya (T: S. Ilyin)	76,924	*Pale Fire* (1962) (O)	85,164
Ada (T: S. Ilyin)	153,621	*Ada or Ardor: A Family Chronicle* (1969) (O)	181,346
Prozrachnyye veshchi (T: S. Ilyin)	23,852	*Transparent Things* (1972) (O)	29,073
Smotri na arlekinov! (T: S. Ilyin)	58,037	*Look at the Harlequins!* (1974) (O)	71,327
Russian Total	1,014,905	*English Total*	1,243,033

3 Unsupervised Classifier

An unsupervised classifier determines patterns in the data without making use of assigned labels, hence it can be considered a *more objective* method, the differences (if) discovered are more pronounced and generally, if labels are provided, a clustering result can be easily reproduced by a supervised classifier.

Nabokov's works can be regarded from multiple perspectives of linguistic phenomena which might go beyond the two languages that we consider here - Russian and English - possibly including French and other languages that the author might have had contact with. Therefore, we choose not to use a label-based supervised

classifier to avoid having any prior expectation of the results. Our method is based on distance similarities between vector representations of documents, so the results are determined by the features and the similarity measure considered.

The classifier is based on a generalization of Ward's method [34] developed initially by Szekely and Rizzo [31] with a restriction for Euclidean distances. Our preliminary study [27] on a smaller corpus of Nabokov's novels already indicates a compatibility point with Burrows' Delta [2] similarity measure. However, in our previous study we do not provide the theoretical background behind the clustering algorithm in connection with any similarity metric.

The process starts with N clusters for each document and it consecutively merges two clusters at each step based on the minimum e distance. Given two classes $\mathscr{A} = \{A_1, \cdots, A_p\}$ and $\mathscr{B} = \{B_1, \cdots, B_q\}$ containing vector representations of documents, and $D\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ any similarity metric, the linkage criterion has the following mathematical formulation:

$$e^D(\mathscr{A}, \mathscr{B}) = \frac{pq}{p+q}\Big(\frac{2}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} D(A_i, B_j) - \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} D(A_i, A_j) - \frac{1}{q^2}\sum_{i=1}^{q}\sum_{j=1}^{q} D(B_i, B_j)\Big) \tag{1}$$

The Lance-Williams parameters [19,31] for this linkage function are identical with the ones for Ward's method for any positive similarity measure D. A clustering result is usually rendered as a dendrogram - a binary tree in which the documents represent the leaves and the sub-clusters are defined by different subtrees starting from the root. In this paper we consider a cluster to be any part of a dendrogram tree, including the entire dendrogram. Our approach combines the single linkage criterion (by which two classes are merged given the smallest distance or nearest neighbor) with a custom objective function which in our case is the general e^D.

The sequential process of joining two clusters at the minimum e^D distance, induces an ultrametric over the space of documents, for which the triangle inequality has a stronger form: $e^D(\mathscr{A}, \mathscr{B}) < \max\{e^D(\mathscr{A}, \mathscr{C}), e^D(\mathscr{C}, \mathscr{B})\}$. Our approach is consistent with previous studies [3,21] which discuss the fact that single linkage and Ward's method always produce monotonic dendrograms, unlike other linkage criteria like UPGMC or WPGMC [7].

To evaluate the results, we make use of the maximum F_1 measure for each class [30]. For a cluster $\mathscr{C}$ and a class K, the precision (P) and recall (R) are defined as:

$$P(\mathscr{C}, K) = \frac{\#\text{ of elements of class } K \text{ in cluster } \mathscr{C}}{|K|}$$

$$R(\mathscr{C}, K) = \frac{\#\text{ of elements of class } K \text{ in cluster } \mathscr{C}}{|\mathscr{C}|}$$

where $|.|$ denotes the cardinal of a set.

The F_β measure is defined by the following formula:

$$F_\beta(\mathscr{C}, K) = (1 + \beta^2) \cdot \frac{P(\mathscr{C}, K) \cdot R(\mathscr{C}, K)}{(\beta^2 \cdot P(\mathscr{C}, K)) + R(\mathscr{C}, K)}$$

The parameter β is used to adjust the importance of precision and recall. For a hierarchical clustering algorithm the maximum precision is attained for any leaf-cluster while the maximum recall is obtained for the entire dendrogram-cluster. To equally weight precision and recall for each class, we select the maximum corresponding F_1 score.

The value of the F_1 score evaluates the degree of compactness of each class. If a class has elements dispersed in different clusters, the corresponding F measure will have a small value.

4 Ranked Lexical Features

Function words or the closed class words (conjunctions, prepositions, pronouns, determiners and particles) have long been studied for authorship attribution [16,17], proving to be a strong indicator of an author's fingerprint. Dinu et al. [4] used these words to uncover the pastiche of a Romanian writer who convinced the literary critics into believing he had discovered the lacking part of an unfinished novel. In such cases any additional authorship results may change the way an author is perceived, as Foucault [9] points out, the concept of *author* is a social construct which reaches beyond the limits of written texts.

The list of English function words, which we also employ in our study, was used to detect translation vs. original texts by Volansky et al. [33]. For Russian, we have constructed the list of function words with all their declensions by crawling Wiktionary [1] a collaborative on-line resource.

The documents are represented as a vector of ranks corresponding to each feature. Our previous approach on a smaller version of this corpus [27] offered good results as well as other previous studies on pastiche detection [5] or text similarity [28]. The idea is to translate the bag-of-words representation of the documents into a rank-vector representation by replacing the frequencies with their corresponding ranks in the document, such that the most frequent word is assigned rank one, the second most frequent rank two, and so on. We state that the ranks are tied when two or more frequencies are equal, in which case we assign the average between the competing, tied ranks. This type of weighting has its roots in Spearman's rank correlation coefficient which indicates the direction of association between two random variables. Forsyth and Sharoff [8] tested the quality of Spearman's correlation for text similarity demonstrating that the approach outperforms a multitude of standard methods.

Using ranks instead of frequencies on text similarity measurements is a good practice for two main reasons: (1) it reduces the bias arising from documents of different size and (2) all the obtained ranked vectors have the same L^1 norm: $\|X\|_1 := \sum_{i=1}^{n} |x_i|$. Where X is any vector of ranks obtained from the bag-of-words and x_i is the rank value corresponding to feature i. Geometrically,

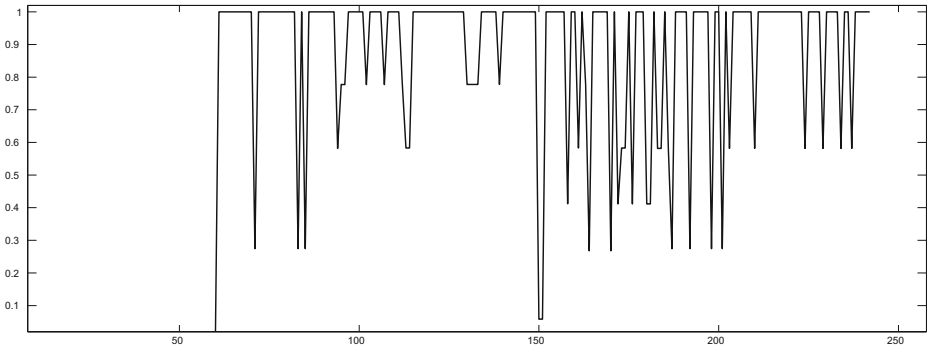

Fig. 1. Plot of the adjusted Rand index between consecutive clusters generated by adding one more word from the list of the first, most frequent function words in the entire English corpus.

the ranked vectors induce an n-dimensional grid, therefore a natural metric to use is the L^1 distance derived from the norm (also called taxicab distance or Manhattan distance):

$$D(X,Y) = \|X - Y\|_1 = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

4.1 Feature Selection

A common problem when carrying text classifications is related to the words that have good discriminative power [18]. In our work, the unsupervised classifier makes use of the pair-wise distances between documents to compute the final dendrogram, therefore, the distances are directly influenced by the features selected [12,20].

First we sort the entire list of function words by their frequency in the entire corpus. Starting from the first 60 function words, we investigate whether changes are produced in the clustering results by using additional features with lower frequencies. To observe the clustering variation, we make use of the adjusted Rand index [13] computed between the "current" and the "previous" cluster.

Given a set of n elements $|S| = n$, and two clustering results $\Psi = \{\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_r\}$ and $\Xi = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_p\}$, construct a contingency table C of r rows and p columns with each value $c_{ij} = |\mathscr{A}_i \cap \mathscr{B}_j|$ being the number of common objects between cluster $\mathscr{A}_i$ and $\mathscr{B}_j$. Let $a_i = \sum_{j=1}^{s} c_{ij}$ be the sum of all the values from the row i and $b_j = \sum_{i=1}^{r} c_{ij}$ all the values from the column j. Then the adjusted Rand index as it is defined by Hubert and Arabie [13] is

$$ARI(\Psi,\Xi) = \frac{\sum_{ij}\binom{c_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{3}$$

If for example we add k consecutive function words for which we obtain identical clusters, then the k features are considered stable and we store them in a unique hash corresponding to the cluster produced. The hash key is obtained from the parenthesis representation of the binary tree, with all the subtrees being sorted lexicographically, for each dendrogram generated sequentially. This way we can have an exploratory technique to account which features contribute to which results. The final feature selection is not necessarily based on a label assignment, but rather it is decided based on the result with the maximal number of features [27]. In Fig. 1 we plot the sequential index values computed for the English corpus. We note that up to the first 150 function words, the clusters are more stable since sequences of different features produce similar results.

5 Similarity Measures and Results

5.1 Manhattan Distance

The most natural measure to be applied in an L^1 space is Manhattan distance. In combination with vectors of ranks it can also be encountered under the name of Spearman's foot-rule or Rank distance [6].

One important property of this metric is its *rank type invariance*: the distance remains unchanged if our tied ranked vectors are obtained by an ascending ordering relation (e.g. assign rank one to the most frequent function word, rank two to the second most frequent and so on) or by a descending ordering relation when rank one is assigned to the most infrequent word and so on. To prove this, we have to observe that for some frequencies $\{f_1 > f_2 > \cdots > f_n\}$, that generated an ascending tied rank $X_> = \{x_1, \cdots, x_n\}$, its descending tied rank can be obtained by the next equation from $X_>$:

$$X_< = (n - X_>) + 1 \tag{4}$$

We observe now that a reverse ranking is produced only with a geometric translation obtained by a subtraction and an addition. Manhattan distance remains unchanged if we translate all the points by the same constant.

This suggests that the use of ranks does not imply just a simple change of the weights, but rather a change of space in which distances between documents become more measurable and more stable (Fig. 2).

5.2 Delta Measure

Delta is a method of measuring stylistic similarities proposed by Burrows [2]. The standard equation for Delta has the following form:

$$\Delta(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{\sigma_i} \right| \tag{5}$$

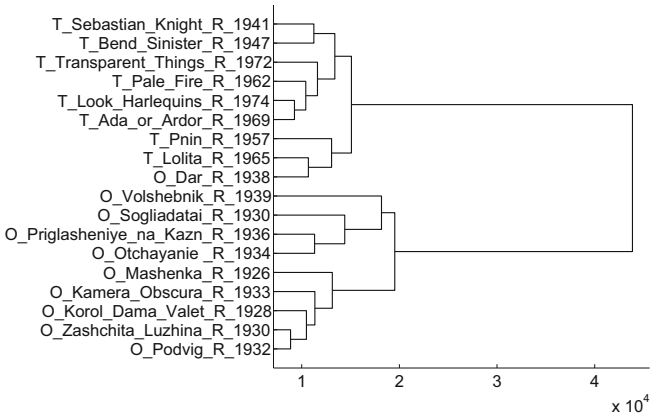

Fig. 2. Result obtained with e^{L^1} linkage criterion using the rankings extracted from the Russian corpus

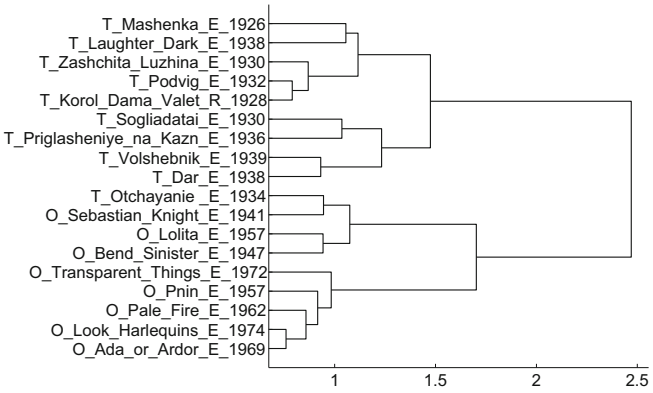

Fig. 3. Result obtained with e^{Δ} linkage criterion using the rankings extracted from the English corpus

where $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_n\}$ are vectors of ranks corresponding to words i and σ_i is the standard deviation of (the rank of) the word i in the given corpus.

Delta is incompatible with the strategy of selecting the entire list of function words due to possible zero standard deviation and other factors discussed by Jockers and Witten [15], in which case a feature selection method becomes almost mandatory. Furthermore, a significant improvement was observed with the use of ranks instead of frequencies. This measure is also invariant to ranking types, the final value of Delta depending on both the ranks of the words within one document and the standard deviation of ranks given all the other documents.

Table 2. Comparison of the F_1 evaluation scores for the entire documents and the texts split into 2000 tokens per chunk and the entire un-split documents.

		L^1		Delta	
		Rank	Freq	Rank	Freq
Russian	O	**0.92**	0.67	0.86	0.88
2000	T	**0.91**	0.65	0.82	0.88
Russian	O	0.94	0.72	**0.94**	0.75
	T	0.94	0.66	**0.94**	0.77
English	O	0.64	0.64	0.64	**0.77**
2000	T	0.69	0.69	0.69	0.69
English	O	0.94	0.73	**0.94**	0.63
	T	0.94	0.64	**0.94**	0.72

A cluster obtained from Delta applied on the English corpus is illustrated in Fig. 3.

Equation 5 is derived from Manhattan distance applied on z-scores of words. For a word i in a given corpus its z-score has the value $z(x_i) = \frac{x_i - \mu_i}{\sigma_i}$ where μ_i is the mean of frequencies x_i of word i. In this case we have the following L^1-like expression for delta measure between two documents X and Y:

$$\Delta(X, Y) = \frac{1}{n} \sum_{i=1}^{n} |z(x_i) - z(y_i)| \tag{6}$$

5.3 F_1 Score

In order to evaluate the F_1 measure [30], we split the each document into smaller chunks of 2000 tokens. For every novel, we randomly extract the same number of chunks so that O and T are not biased by the presence of the larger novels. Moreover, since *Sogliadatai* and *Volshebnik* are considerably smaller in size, we decided to discard them from this analysis.

The F_1 scores obtained in Table 2 indicate that both L^1 and Delta are comparable in terms of results when the full documents are used since, in this scenario, the standard deviation does not have a large impact over the measured similarities. What is more, the use of ranks seems to greatly influence the compactness of the clusters (0.94) while standard frequencies barely score an F-measure above 0.7.

However, when the documents are split into chunks, we observe a significant drop in F-measure regardless of the ranking process. If for Russian the clusters are still quite compact - 0.92 for L^1 with ranks and 0.88 for Delta with frequencies, for English the best score (0.77) is obtained by using standard Delta in combination with frequencies.

We believe there are two causes for this behavior: (1) the chunks are smaller and the features that differentiate the author from translator are less frequent,

making the distance-based similarities less prominent and (2) Nabokov's personal involvement in the translations from Russian may have determined his authorial *fingerprint* to be actively present in the English translations, thus creating a stronger resemblance between translation and original. A close inspection of the large dendrogram resulted shows that the translated English chunks are not homogeneous, but rather spread across different clusters of O. To conclude, interference seems to be more present in the Russian translations over which the author had a minimal contribution.

6 Conclusions

We propose an extension of Ward's hierarchical clustering method that is able to operate with custom user-defined objective functions that are not required to be metrics. Given the consistent results on two different languages, our combination of exploratory methods can be considered reliable for measuring distances between different text documents. Furthermore, our results indicate that ranks do improve the evaluation F-scores when the number of training examples is small. Both the L^1 metric and Delta are rank type invariant, which means the results are identical if we assign rank one to the most frequent feature and so on, or rank one to the most infrequent feature and so on.

Our adapted clustering algorithm was able to successfully distinguish between Nabokov's original novels and translations on two different languages with multiple translators involved. Compared to previous work investigating translation [14,33], our results further bring into discussion the influence of the author over the translation and a possible link between interference and language transfer. Hence, we show that it is difficult to correctly classify between author and author-as-translator, especially when the size of the documents is small and when a possible imprint of language transfer could influence the overall results. Translations highly depend on the choices a translator makes to reproduce the initial style of the text, but these decisions further depend on the O vs. T linguistic and cultural differences.

Last but not least, we further add a proof to the fact that the *fingerprint* of an author can be revealed by his use of function words, *fingerprint* which can get masked under the effect of translation.

References

1. Wiktionary. ru.wiktionary.org. Accessed in June 2013
2. Burrows, J.F.: Delta: a measure of stylistic difference and a guide to likely authorship. Literary Linguist. Comput. **17**(1), 267–287 (2002)
3. Carlsson, G.E., Mémoli, F.: Characterization, stability and convergence of hierarchical clustering methods. J. Mach. Learn. Res. **11**, 1425–1470 (2010)
4. Dinu, L.P., Niculae, V., Şulea, O.M.: 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, pp. 72–77. Association for Computational Linguistics, Stroudsburg (2012)

5. Dinu, L.P., Niculae, V., Şulea, O.M.: 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, pp. 72–77 (2012)
6. Dinu, L.P., Popescu, M.: Comparing statistical similarity measures for stylistic multivariate analysis. In: RANLP, pp. 349–354. Association for Computational Linguistics, Borovets (2009)
7. Everit, B., Landau, S., Leese, M.: Cluster Analysis. Hodder, London (2001)
8. Forsyth, R., Sharoff, S.: Document dissimilarity within and across languages: a benchmarking study. Literary Linguist. Comput. **29**, 6–22 (2014)
9. Foucault, M.: What Is an Author?. State University Press of New York, Albany (1987)
10. Gorski, B.: Nabokov vs. Набоков: A literary investigation of linguistic relativity. Vestnik, J. Russ. Asian Stud. (8) 56–78 (2010) http://www.sras.org/nabokov_vs_nabokov_linguistic_relativity
11. Hallé, P.A., Best, C.T., Levitt, A.: Phonetic vs. phonological influences on french listeners' perception of american english approximants. J. Phonetics **27**(3), 281–306 (1999)
12. Hoover, D.L.: Testing burrows's delta. Literary Linguist. Comput. **19**(4), 453–475 (2004)
13. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
14. Ilisei, I., Inkpen, D., Corpas Pastor, G., Mitkov, R.: Identification of translationese: a machine learning approach. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 503–511. Springer, Heidelberg (2010)
15. Jockers, M.L., Witten, D.M.: A comparative study of machine learning methods for authorship attribution. Literary Linguist. Comput. **25**, 215–223 (2012)
16. Juola, P.: Authorship attribution. Found. Trends Inf. Retrieval **1**(3), 233–334 (2006)
17. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. **60**(1), 9–26 (2009)
18. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: unmasking pseudonymous authors. J. Mach. Learn. Res. **8**, 1261–1276 (2007)
19. Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies: 1. hierarchical systems. Comput. J. **9**(4), 373–380 (1967)
20. Madigan, D., Genkin, A., Lewis, D.D., Lewis, E.G.D.D., Argamon, S., Fradkin, D., Ye, L., Consulting, D.D.L.: Author identification on the large scale. In: Proceedings of the Meeting of the Classification Society of North America (2005)
21. Milligan, G.W.: Ultrametric hierarchical clustering algorithms. PSYCHOMETRIKA **44**(3), 343–346 (1979)
22. Mosteller, F., Wallace, L.D.: Inference in an authorship problem. J. Am. Stat. Assoc. **58**(302), 275–309 (1963)
23. Nabokov, V.: Lolita. Penguin Books Limited, UK (2012)
24. Nabokov, V.: Speak, Memory: An Autobiography Revisited. Vintage International, New York (1989)
25. Nabokov, V.: Eugene Onegin. A Translation from the Russian of Aleksandr Pushkin's (1833) Yevgeniy Onegin (1990)
26. Nisioi, S.: Feature analysis for native language identification. In: Gelbukh, A. (ed.) CICLing 2015, Part I. LNCS, vol. 9041, pp. 644–657. Springer, Heidelberg (2015)
27. Nisioi, S., Dinu, L.P.: A clustering approach for translationese identification. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, pp. 532–538, September 2013

28. Popescu, M., Dinu, L.P.: Comparing statistical similarity measures for stylistic multivariate analysis. In: RANLP 2009 Organising Committee/ACL RANLP, pp. 349–354 (2009)
29. Selinker, L., Rutherford, W.: Rediscovering Interlanguage. Applied Linguistics and Language Study. Routledge, New York (2014)
30. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques (2000)
31. Szekely, G.J., Rizzo, M.L.: Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. J. Classif. **22**, 151–183 (2005)
32. Tsvetkov, Y., Twitto, N., Schneider, N., Ordan, N., Faruqui, M., Chahuneau, V., Wintner, S., Dyer, C.: Identifying the l1 of non-native writers: the cmu-haifa system. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 279–287. Association for Computational Linguistics, Atlanta, June 2013
33. Volansky, V., Ordan, N., Wintner, S.: On the features of translationese. Digit. Scholars. Humanit. **30**(1) 98–118 (2015) doi:10.1093/llc/fqt031
34. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **301**(58), 236–244 (1963)