DHQ: Digital Humanities Quarterly
2021
Volume 15 Number 1

Can an author style be unveiled through word distribution?

Giulia Benotto <giulia_dot_benotto_at_extrasys_dot_it>, Extra Group

Abstract

The inclusion of semantic features in the stylometric analysis of literary texts appears to be poorly investigated. In this work, we experiment with the application of Distributional Semantics to a corpus of Italian literature to test if words distribution can convey stylistic cues. To verify our hypothesis, we have set up an Authorship Attribution experiment. Indeed, the results we have obtained suggest that the style of an author can reveal itself through words distribution too.

Introduction

Stylometry, that is the application of the study of linguistic style, offers a means of capturing the elusive character of an author's style by quantifying some of its features. The basic stylometric assumption is that each writer has a "human stylome" [Van Halteren et al. 2005], that is a set of certain stylistic idiosyncrasies that define their style. Analysis based on stylometry is often used for Authorship Attribution (AA) tasks, since the main idea behind computationally supported AA is that by measuring some textual features, it is possible to distinguish between texts written by different authors [Stamatatos 2009]. One of the less investigated stylistic features is the way in which authors use words from a semantic point of view, e.g. if they tend to use more, when dealing with polysemous words, a certain sense over the others, or senses that differ (even slightly) from the one that is more commonly used (as it happens, typically, in poetry). It would then be interesting to discover if the semantics an author bestows to words is actually part of its "human stylome." Computing semantics, though, is far from easy. 1

A possible approach to the analysis of this characteristic is to consider the textual contexts in which certain words appear. According to Distributional Semantics (DS), certain aspects of the meaning of lexical expressions depend on the distributional properties of such expressions, or better, on the contexts in which they are observed [Lenci 2008] [Miller and Charles 1991]. The semantic properties of a word can then be defined by inspecting a significant number of linguistic contexts, representative of the distributional behavior of such word. 2

In this work, we would like to investigate if the analysis of the distribution of words in a text can be exploited to provide a stylistic cue. In order to inspect that, we have experimented with the application of DS to the stylometric analysis of literary texts belonging to a corpus constituted by texts pertaining to the work of six Italian writers of the late nineteenth century. In the following, Section 2 both provides a brief survey on the works related to stylometry and introduces the fundamental Distributional Semantics concepts on which this works relies upon. Section 3 describes the approach together with the corpus used to conduct our investigation and Section 4 discusses results. Finally, Section 5 draws some conclusions and outlines some possible future works. 3

Background Knowledge

Stylometry

Even if the first attempt at computing the writing style of an author dates back to the first half of the 20th century ([Yule 1938] [Yule 1944] [Zipf 1932]), the work that is believed to be the starting point of the so-called "non-traditional" Authorship Attribution is a study by Mosteller and Wallace (1964). They conducted an investigation on the authorship of 4 the “Federalist Paper,” a series of political essays written by John Jay, Alexander Hamilton, and James Madison, 12 of which have ambiguous paternity, being claimed by both Hamilton and Madison. From then on (to, at least, the end of the 1990s), research in AA mainly coincided with “stylometry,” i.e. defining features to quantify the style of an author by using measures based on counting frequencies of words, characters, and sentences [Holmes 1994] [Holmes 1998].

Despite their working well, these systems followed a methodology that underwent some limitations, such as the little statistical homogeneity of the analyzed texts or the fact that the evaluation of the developed methods was mainly intuitive, using corpora that were not controlled for topic; moreover the lack of benchmark data made it difficult to compare different methods. These problems were partially overcome at the end of the 1990s, when the internet made a vast amount of electronic texts available, also highlighting all different areas in which AA could be useful, beyond that of literary research (i.e. intelligence [Abbasi and Chen 2005], criminal and civil law [Chaski 2005] [Grant 2007], computer forensic [Frantzeskou et al. 2006] and so on). Nowadays, the main emphasis on AA tasks regards the objective evaluation of the proposed methods using common benchmark corpora [Juola 2004]. 5

As previously stated, the very first attempts to analyze the style of an author were mainly based on lexical features, such as sentence length count or words length count, which can be applied to any language and corpus without additional requirements [Koppel and Schler 2004] [Stamatatos 2006] [Zhao and Zobel 2005] [Argamon et al. 2007]. Character measures, too, have been proven to be useful in quantifying the writing style. A text can then be viewed as a sequence of characters on which various measures (such as alphabetic, digit, uppercase and lowercase characters count) can be defined [Zheng et al. 2006] [Grieve 2007] [De Vel et al. 2001]. A more elaborate text representation method is based on the assumption that authors tend to use similar syntactic patterns, so syntactic information is employed, being considered a more reliable authorial fingerprint than lexical information [Gamon et al. 2004] [Stamatatos et al. 2000] [Stamatatos et al. 2001] [Hirst and Feiguina 2007] [Uzuner and Katz 2005]. 6

Very few attempts to exploit high-level features for stylometric purposes have been made, due to the fact that tasks such as full syntactic parsing, semantic analysis, or pragmatic analysis cannot yet be handled adequately by current NLP technologies. The most important method of exploiting semantic information, so far, was based on the theory of Systemic Functional Grammar (SFG) [Halliday 1994] and consisted on the definition of a set of functional features that associate certain words or phrases with semantic information, as described in Argamon (2007). 7

However, the goal of our work is to use only information about words distribution, in order to discover if a correlation between an author's style and words distribution exists. In order to analyze words distribution, we rely on the Distributional Hypothesis and, consequently, on Distributional Semantics. Their theoretical basis will be presented in the next subsection. 8

Distributional Semantics

The assumption behind all distributional semantics models (DSMs) is that the notion of semantic similarity can be defined in terms of linguistic distributions. This is known as the Distributional Hypothesis, which is stated as follows: “The degree of semantic similarity between two linguistic expressions a and b depends on the similarity of the linguistic contexts in which a and b can appear.” In accord with this definition, certain aspects of the meaning of lexical expressions depend on the contexts in which they are observed. The semantic properties of a word can then be defined by inspecting a significant number of linguistic contexts, representative of the distributional behavior of such word. 9

Despite the huge consensus reached lately by this theory in Computational Linguistics, its origins reside in the context of Zellig Harris’ proposal of Distributional Semantics as the bedrock of linguistics as a scientific discipline [Harris 1970]. Harris’ proposal was conceived for phonemic analysis and later turned into a general methodology to be applied at every linguistic level. The distribution procedure was regarded as a way for linguists to give a methodological base to their analysis. He then, not only extended his theory to meaning but assumed that meaning could actually be explained on distributional grounds. The Distributional Hypothesis can be considered a cognitive hypothesis about the form and origin of semantic representations [Miller and Charles 1991]. Some of the most influential models for distributional semantics, such as Latent Semantic Analysis (LSA; [Landauer and Dumais 1997]) and Hyperspace Analogue to Language (HAL; [Burgess and Lund 1997]) have been developed for cognitive and psychological research, meant to 10 represent how semantic representations are learned by extracting co-occurrence patterns [Landauer 2007].

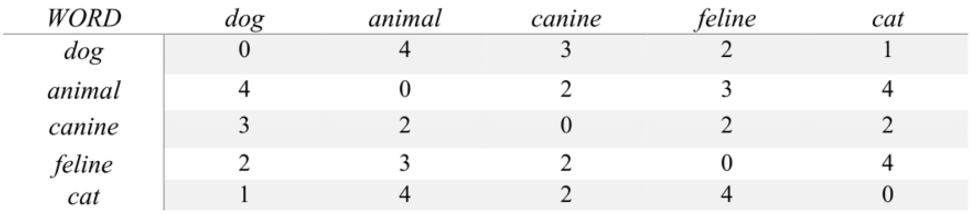
11 Within the corpus linguistics tradition, there was no need to motivate the Distributional Hypothesis as a methodological principle for semantic analysis. This is better summarized in the well-mentioned slogan by Firth: “You shall know a word by the company it keeps” [Firth 1957]. In corpus linguistics, the Distributional Hypothesis is often claimed to be the only possible source of evidence for the exploration of meaning.

12 Distributional Semantics has gained popularity in computational linguistics starting from the late 1980s when there was the progressive predominance of corpus-based statistical methods for language processing. Within this new paradigm, statistical methods were naturally applied to the lexical-semantic analysis. Corpora are indeed connected to Distributional Semantics since they can be used as repositories of linguistic usages, then representing the primary source of information to identify the world distributional properties. Their role has been enhanced by the current availability of a huge collection of texts, contextually with increasingly sophisticated computational linguistics techniques to process them and extract the relevant context feature to build distributional semantic representations. Despite its currently being corpus-based, distributional semantics is not prevented to underline aspects of the format and origin of word meaning and the issue of how and to what extent features extracted from the linguistic input actually shape meaning.

13 The way in which it is possible to proceed in order to infer a geometric representation starting from distributional information can be originated from Harris (1970), who writes that “the distribution of an element will be understood as the sum of all its environments.” In linguistics, an environment is called a context, and we here assume a context to be the setting of a word, phrase, etc., among the surrounding words, phrases, etc., often used for helping to explain the meaning of the word, phrase, etc.

14 One way to collect this information is to tabulate the contextual information, so that for each word we provide a list of the co-occurrents of the word and the number of times they have co-occurred. In a second step, we take away the actual words and only leave the co-occurrence counts. Also, we make each list equally long by adding zeroes in the places where we lack co-occurrence information. We also sort each list so that the co-occurrence counts for each context come in the same places in the lists. The mathematical backbone of Latent Semantic Analysis [Landauer and Dumais 1997], is Singular Value Decomposition, a well-known linear algebra technique aimed at extracting the most informative dimensions in a matrix of data and here used to reconstruct the latent structure behind the distributional hypothesis [Deerwester et al. 1990]. The names vector semantics, word or semantic spaces merely highlight specific mathematical techniques used to formalize the notion of contextual representation. This information can then be modeled as vectors, as described in Schütze (1992), Schütze (1993), who builds context vectors (which he calls “term vectors” or “word vectors”) in the following way: co-occurrence counts are collected in a words-by-words matrix, in which the elements record the number of times two words co-occur within a set window of word tokens.

15 Context vectors are then defined as the rows or the columns of the matrix (the matrix is symmetric, so it does not matter if the rows or the columns are used). A cell fij of the co-occurrence matrix records the frequency of occurrence of the word i in the context of the word j or of the document j, as shown in Figure 1.



WORD	*dog*	*animal*	*canine*	*feline*	*cat*
dog	0	4	3	2	1
animal	4	0	2	3	4
canine	3	2	0	2	2
feline	2	3	2	0	4
cat	1	4	2	4	0

Figure 1. Words-by-words co-occurrences matrix.

Context vectors do not only allow us to go from distributional information to a geometric representation, but they also make it possible for us to compute proximity between words. Thus, the point of the context vectors is that they allow us to define (distributional, semantic) similarity between words in terms of vector similarity. There are many ways to compute the similarity between vectors, and the measures can be divided into similarity measures and distance measures. The difference is that similarity measures produce a high score for similar objects, whereas distance measures produce a low score for the same objects: large similarity equals small distance, and conversely. A convenient way to compute normalized vector similarity is to calculate the cosine of the angles between two vectors x and y, defined as:

$$sim_{cos(\vec{x},\vec{y})} = \frac{x \cdot y}{|x| \cdot |y|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

The cosine measure corresponds to taking the scalar product of the vectors and then dividing by their norms. It is the most frequently utilized similarity metric in word-space research. It is attractive because it provides a fixed measure of similarity: it ranges from 1 for identical vectors, over 0 for orthogonal vectors. Figure 2 shows an example of the graphical representation of words vectors related to the matrix depicted in Figure 1. The words “cat” and “dog” in the matrix above are depicted as never appearing together in the same context. This does not mean the ey are not semantically similar, because they can actually happen in really similar contexts, it just means they don't appear together in the context windows we defined. Anyway, here, we represent “cat” and “dog” as dissimilar, being the angle between their vector of almost 90 degrees. The vector representative of the word “animal” is instead as similar to that of “dog” than to that of “cat”, while the vector representative of the word “canine” is closer to the vector representative of “dog” than to “cat” and the vector representative of “feline” is closer to “cat” than to “dog” meaning that “canine” is more semantically similar to “dog” and “feline” is more semantically similar to “cat”. Also, the vectors of “canine” and “feline” are both close to “animal”, suggesting that the two words are often used in similar contexts in the texts analyzed to generate the co-occurrence vectors here represented. Of course, this example is not representative of the linguistic and semantic reality of things, but entirely indicative and apt to properly describe and illustrate the concept of vectorial representation of semantic similarity.

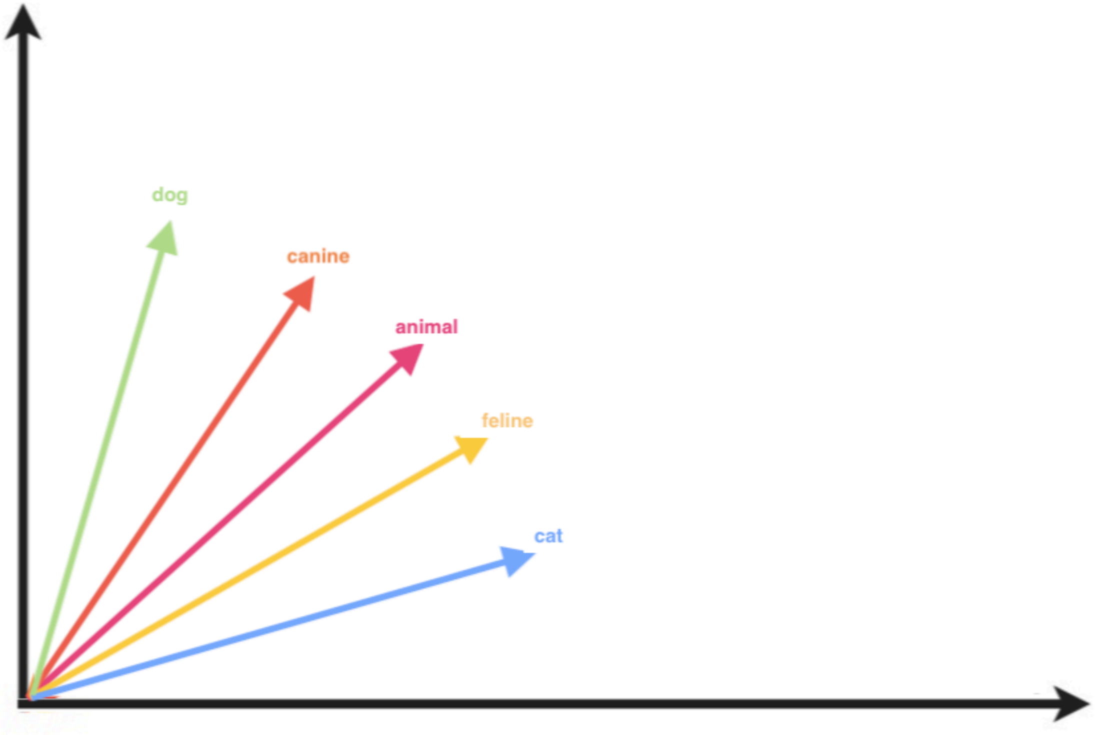


Figure 2. Vector Similarity.

Experimental Setup

First, we want to specify that it is not our purpose to propose new ways to improve state-of-the-art AA algorithms. Indeed, our aim is just to verify the hypothesis that the distribution of words can provide an indication of a distributional stylistic fingerprint of an author. To do this, we have set up a simple classification task. Subsection 3.1 briefly depicts the data set we used, Section 3.2 describes why and how we chose the authors that would constitute our dataset and Section 3.3 depicts the steps implemented in our experiment. 16

Data Set Construction

In order to build the reference and test corpora, we started from texts pertaining to the work of six Italian writers working at the turn of the 20th century, namely, Luigi Capuana, Federico De Roberto, Luigi Pirandello, Italo Svevo, Federigo Tozzi and Giovanni Verga. We chose contiguous authors in a chronological sense, whose texts are available in digital format (in fact we could not do a similar survey on the narrative of the 1990s because it is still under copyrights). Indeed, we used texts freely available for download from the digital library of the Manunzio project, via the LiberLiber website (www.liberliber.it). Since they were encoded in various formats, such as .epub, .odt, and .txt, our pre-processing consisted in converting them all in .txt format and getting rid of all XML tags, together with footnotes and editors' notes and comments. 17

Authors and Texts Choice

In between the 1875 and the early 1900s, a literary movement peaked in Italy: Verismo (meaning "realism", from Italian vero, meaning "true"). The main exponents of this literary movement, as well as the authors of its manifesto, were Giovanni Verga and Luigi Capuana. Verismo did not constitute a formal school, but it was still based on specific principles, its birth being influenced by a positivist climate which put absolute faith in science, empiricism, and research and which developed from 1830 until the end of the 19th century. 18

All the authors selected to build the corpus used for this work pertained to the temporal span in which the Literary Verismo developed, but not all of them are proponents of such genre. Indeed, three of the selected authors are considered to be Verist Authors (Giovanni Verga, Federico De Roberto, and Luigi Capuana) while three (Luigi Pirandello, Federico Tozzi, and Italo Svevo) are representative of another Literary Movement: Modernism. We decided to choose texts pertaining to those authors and literary movements firstly because of their being all written in the same temporal span. This allowed us to get rid of any eventual lexical bias, due to the difference in languages of works published in different epochs. 19

Also, the selection was performed having in mind an eventual future evolution of this work. Using texts pertaining to the same period, but to different literary movement, would allow us to investigate the ability of our method in recognizing the literary movement the texts pertain to, instead of the author that wrote them. This style-based text categorization tasks, known as genre detection, is quite similar to authorship attribution, even if there are characteristics that distinguish the one from the other. An important question to investigate, then, is how it would be possible to discriminate between two basic factors: authorship and genre, and if semantics could be useful not only for recognizing the author of a literary work but also the literary genre it pertains to. 20

Another line of research that has not been adequately examined so far is the development of robust attribution techniques that can be trained on texts from one genre and applied to texts of another genre by the same authors. The way we selected and balanced the texts composing the corpus could be useful for this task, too. 21

Experiment Description

According to Rudman (1997), a striking problem in stylometry is due to the lack of homogeneity of the examined corpora, in particular to the improper selection or fragmentation of the texts that might cause alterations in the writers' style. As already depicted in the previous section, the corpus has been built according to this assumption, trying to use texts pertaining to the same time span and balanced between the two selected literary movements. Also, in order to 22 create balanced reference corpora, i.e. covering all the authors' different stylistic and thematic phases, for each author, as shown in Figure 3, we built a reference corpus as the composition of the 70% of every single work (usually a novel). The same technique was used to create the test corpus by using the remaining 30% of each work. Typical AA approaches consist of analyzing known authors and assigning authorship to previously unseen text on the basis of various features. Train and test sets should then contain different texts. Contrary to the classical AA task, our train and test sets contain different parts of the same texts. Indeed, with this experiment, we wanted to understand if the semantics that an author bestows to a word, is peculiar to his writing. To prove this, we wanted to cover all the different stylistic and thematic phases an author can go through during his activity, hence the partition of all his texts in a reference and a test portion.

Our work relies on the assumption that the works of an author are representative of the author's thought, so it is assumed that the semantics that an author associates with certain words are representative of its thoughts. One possible flaw of this kind of approach is that if an author changes the semantics that he associates with concepts in different works overtime, the method might not work. It can happen that an author who has a long career changes its point of view on things and this should obviously reflect on its works. This is one of the main reasons why we decided to use all the works from each author as training text. We wanted to take into account each different phase an author may go through during its career, especially considering changes in the semantics associated with concepts. 23

Using all the works of each author allows us to have complete photography of the author itself, and allows to understand the semantics associated with concepts through all its work, even accounting for changes. In fact, the association between words extracted using our method would highlight changes in semantics by changing the score associated with a pair of words. Let's hypothesize that a strong association for the young Verga (i.e. for Verga in his first works) is sun and joy, while later on the strongest association is, let's say, moon and joy. Our hypothesis would be that, while in first works Verga associated the concept of "sun" with that of "joy," later on, is the concept of "moon" that is associated with that of "joy." Deciding to use some works of Verga as train as some other as tests might then be deceiving, because what is semantically true for his first work is not true later on. Using our method, if an association is true just for some works, and not for all its score is evened out and the pair of words is not that semantically relevant, and thus is not used for classification. Pair with high score are those for which the association between the concepts are true throughout all the work of an author, or reports a score that is so high in one or more work, that is could not be evened out from the score reported in all the other association from all the other works from that particular author. 24

We then analyzed each reference and test corpora with a Part-of-Speech (PoS) tagger and a lemmatizer for Italian [Dell'Orletta et al. 2014]. For every author, we built two lists of word pairs (with their lemma and PoS), one relative to the tagged reference corpus (reference pairs) and the other to the tagged test set (test pairs), where each word was paired with all the other words with the same PoS. We also filtered the pairs to leave only nouns, adjectives and verbs. Starting from the tagged corpora, we built two words-by-words matrixes of co-occurrence counts for each author. Being the corpus relatively small and not having particular computability issues, we chose not to apply decomposition techniques to reduce the size of the matrixes (and thus not losing any information). We performed different empiric setup of the window's size and chose the one that showed more suitable results, according to what is stated by Kruszewski and Baroni: the context window was then set to 3 words prior and 3 words following the one under examination [Kruszewski and Baroni 2014]. The chosen DS model [Baroni and Lenci 2010] was applied to each matrix to calculate the cosine between the vectors representing the two words of each pair. This allowed us to evaluate the semantic relatedness between the words by assessing their proximity in the distributional space as represented by the cosine value: as explained in Section 2.2, the more this value tends to 1, the more the two words of the pair are considered to be related. We then obtained two related word pair (RWP) lists for each author A: RWPrefA and RWPtestA. Figure 3 depicts the process described above. 25

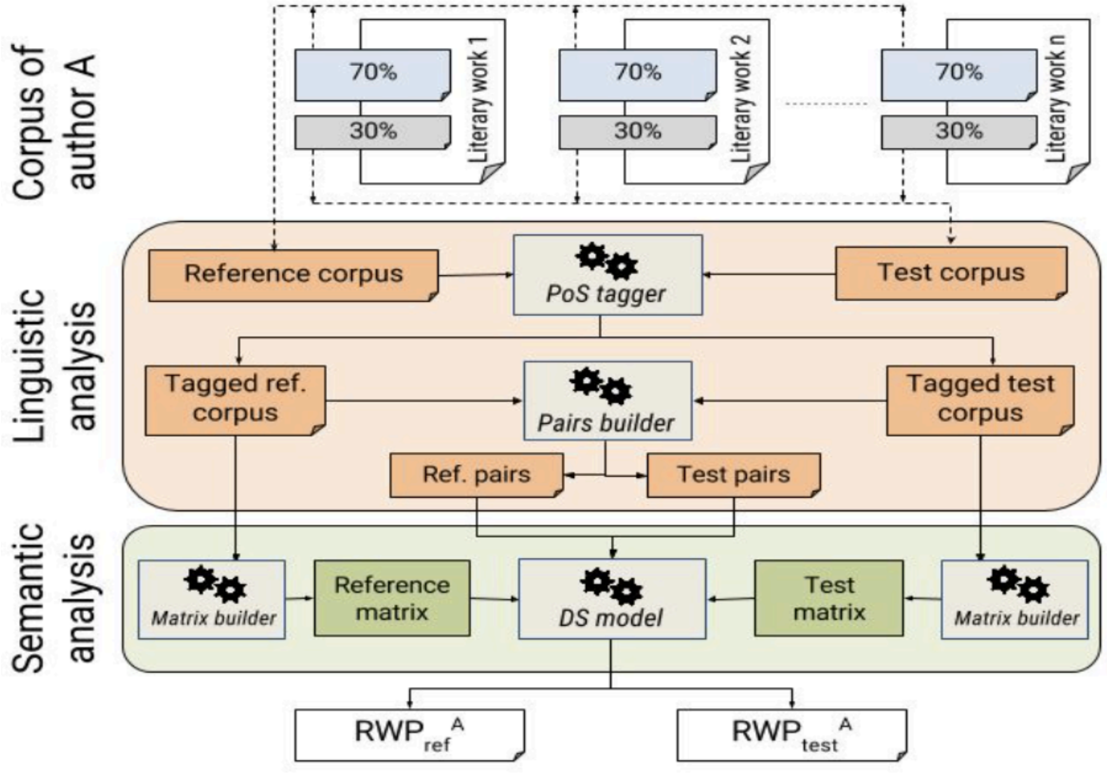
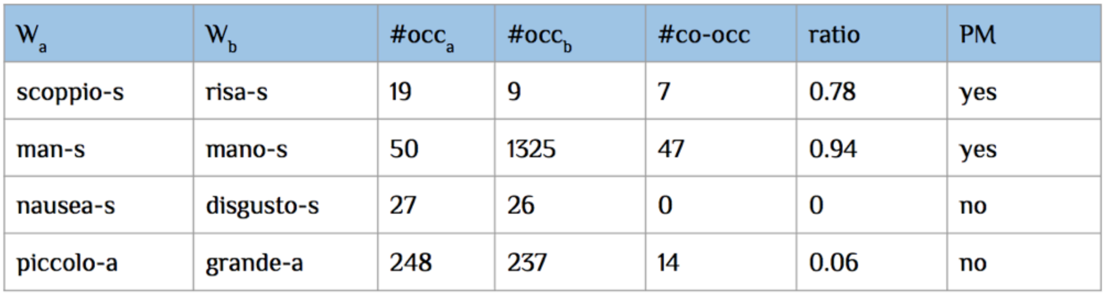


Figure 3. RWPref and RWPtest creation process for an author.

Hypothesis and Discussion

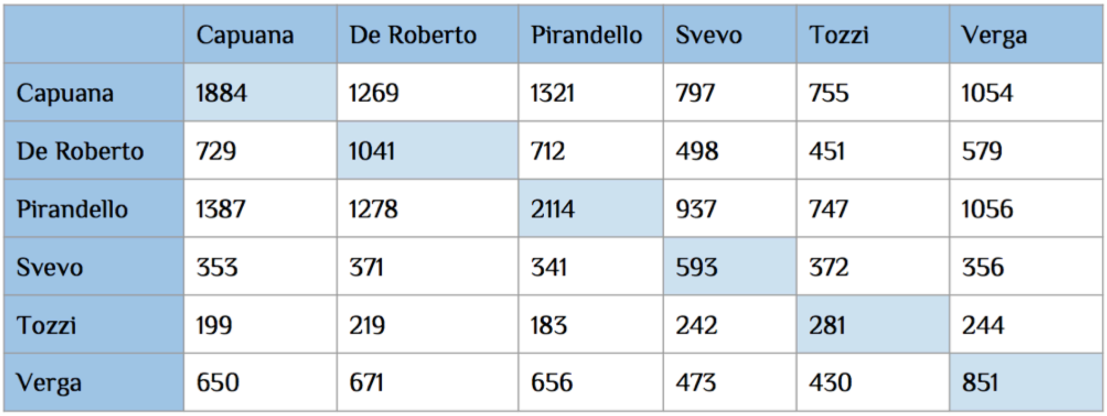
Since we wanted to focus on the analysis of the semantic distribution of words, we decided to exclude any possible “lexical bias.” For this reason, we restricted the analysis on a common vocabulary, i.e. a vocabulary constituted by the intersection of the six authors’ vocabularies. In this way, we prevent our classifier to exploit, as a feature, the presence of words used by some (but not all) of the authors. Moreover, we removed from the RWPtest lists all those pairs of words occurring frequently together in the same context, since they might constitute a multiword expression that, once again, could be pertaining with the signature lexicon of each author. To remove them, we computed the number of times (#co-occ in Figure 4) they appeared together in the context window, as well as their total number of occurrences (#occa and #occb) and we excluded from the analysis those pairs for which the ratio between the number of co-occurrences and the total occurrences of the less frequent word was higher than the empirically set threshold of 0.5. The first two pairs of Figure 4 would be removed as probable multiword (PM column in Figure 4): “scoppio” (burst) and “risa” (laughter) could mostly co-occur in “scoppio di risa” (meaning “burst of laughter”) and the words “man” and “mano” (both meaning “hand”) could mostly co-occur in “man mano” (meaning “little by little,” or “progressively”).



W_a	W_b	$\#occ_a$	$\#occ_b$	#co-occ	ratio	PM
scoppio-s	risa-s	19	9	7	0.78	yes
man-s	mano-s	50	1325	47	0.94	yes
nausea-s	disgusto-s	27	26	0	0	no
piccolo-a	grande-a	248	237	14	0.06	no

Figure 4. An example of co-occurring RWPs from Pirandello’s test list: the first two pairs would be removed.

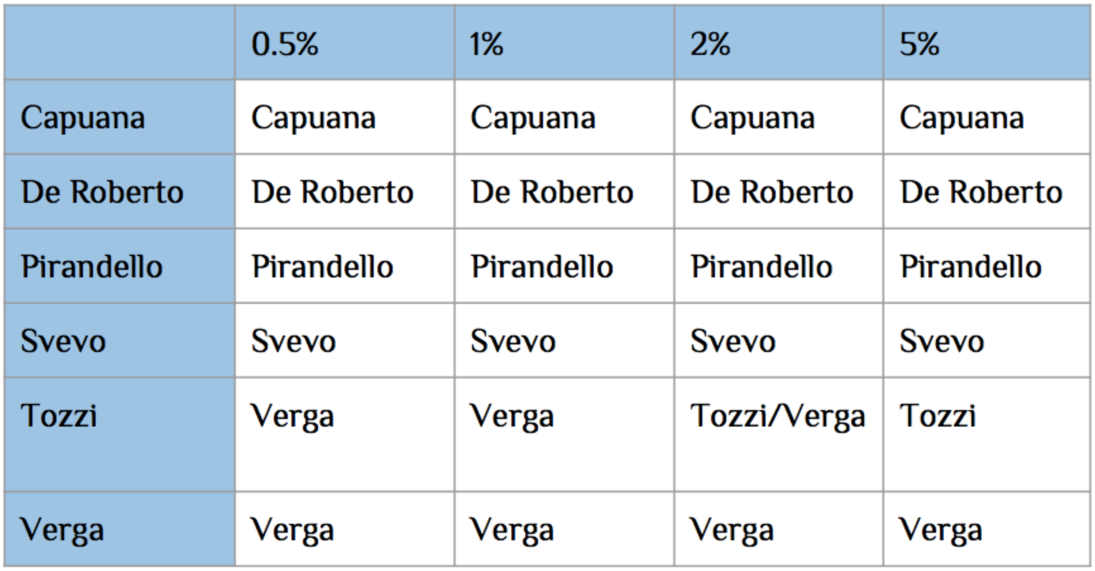
Finally, we reduced the size of the six RWPref and RWPtest lists by sorting them in decreasing order of the cosine value and then by keeping the pairs with the highest cosine, selected using a percentage parameter θ as a threshold. We chose to introduce the parameter θ for two reasons: first of all we wanted to avoid the classification algorithm to be disturbed by noisy (i.e. not significative) pairs which would not hold any relevant stylistic cue, also, we would like to ease a literary scholar in the interpretation of the results by having to analyze just a limited selection of (potentially) semantically related word pairs. For the last phase of our experiment, we defined a classification algorithm to test the effective presence of stylistic cues inside the obtained RWPtest lists. We defined a classifier using a nearest-cosine method to attribute each test list to an author. The method consisted in searching for a pair of words contained in the test list inside each reference list and incrementing by 1 the score of the author whose reference list included the pair with the more similar cosine value (i.e. having the minimum difference): the chosen author was the one with the highest score. Figure 5 shows the classification results for θ = 5%.



	Capuana	De Roberto	Pirandello	Svevo	Tozzi	Verga
Capuana	1884	1269	1321	797	755	1054
De Roberto	729	1041	712	498	451	579
Pirandello	1387	1278	2114	937	747	1056
Svevo	353	371	341	593	372	356
Tozzi	199	219	183	242	281	244
Verga	650	671	656	473	430	851

Figure 5. Classification results, obtained via the nearest-cosine method for θ = 5%.

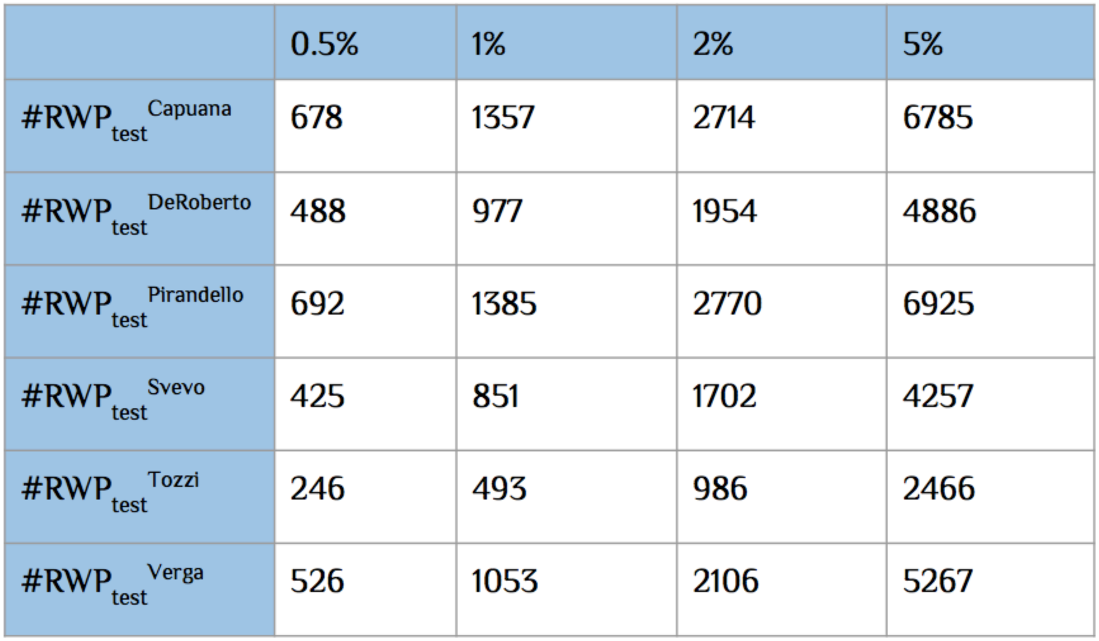
As summarized in Figure 6, the correct classification of all RWPs in RWPtest lists has been obtained with a θ value of 5%.



	0.5%	1%	2%	5%
Capuana	Capuana	Capuana	Capuana	Capuana
De Roberto	De Roberto	De Roberto	De Roberto	De Roberto
Pirandello	Pirandello	Pirandello	Pirandello	Pirandello
Svevo	Svevo	Svevo	Svevo	Svevo
Tozzi	Verga	Verga	Tozzi/Verga	Tozzi
Verga	Verga	Verga	Verga	Verga

Figure 6. Results of the classification. Classification errors are highlighted.

To help in interpreting the failure of the algorithm in classifying Tozzi’s test list for θ values lower than 5% (as shown in Figure 6) we calculated the cardinality of the RWPtest lists for each author with the change in θ value (Figure 7).



	0.5%	1%	2%	5%
#RWP$_{test}^{Capuana}$	678	1357	2714	6785
#RWP$_{test}^{DeRoberto}$	488	977	1954	4886
#RWP$_{test}^{Pirandello}$	692	1385	2770	6925
#RWP$_{test}^{Svevo}$	425	851	1702	4257
#RWP$_{test}^{Tozzi}$	246	493	986	2466
#RWP$_{test}^{Verga}$	526	1053	2106	5267

Figure 7. Cardinality of RWPtest for each author and for each θ value.

It is possible to observe how the choice of θ influences the correct classification of Tozzi's test list. Indeed, the use of a θ sense below 5% has the effect of remarkably reducing an already small test list (RWPtextTozzi) as shown in Figure 7. It is apparent that increasing the value of θ and consequently the number of significant RWPs that are analyzed, the system is able to correctly classify RWPtestTozzi (see the values in Tozzi's row of Figure 6).

Conclusion and Next Steps

In this paper, we investigated the possibility that an analysis of the semantic distribution of words in a text can be potentially exploited to get cues about the style of an author. In order to validate our hypothesis, we conducted the first experiment on six different Italian authors. Of course, it is not our intent, with this paper, to define new methods for enhancing state-of-the-art authorship attribution algorithms. However, the obtained results seem to suggest that the way words are distributed across a text, can provide a valid stylistic cue to distinguish an author's work. In light of what we have shown up to this point, the direction of our next steps can be twofold. On the one hand, our research will focus on detecting and providing useful indications about the style of an author. This can be done by highlighting, for example, atypical distributions of words (e.g. with contrastive methods) or by analyzing their distributional variability. Furthermore, it could be interesting to use a different distributional measure than the cosine, to test our hypothesis. On the other hand, it would be interesting to confront the computational task of authorship attribution, by measuring the effective contribution that a feature based on distributional semantics would provide to a canonical classification process. Also, as highlighted in Section 3.2.3, another interesting development of this work would regard the investigation of the ability of our method in recognizing the literary movement the texts pertain to, instead of the author that wrote them.

Works Cited

Abbasi and Chen 2005 Abbasi A., Chen H. 2005. “Applying authorship analysis to extremist-group web forum messages.” *IEEE Intelligent Systems*, 20(5), 67-75.

Argamon et al. 2007 Argamon S., Whitelaw C., Chase P., Hota S. R., Garg N., and Levitan S. 2007. “Stylistic text classification using functional lexical features.” *Journal of the American Society for Information Science and Technology*, 58(6):802– 822, April.

Baroni and Lenci 2010 Baroni M. and Lenci A. 2010. “Distributional memory: A general framework for corpus-based semantics.” *Computational Linguistics*, 36(4):673–721.

Buitelaar et al. 2014 Buitelaar P., Aggarwal N., and Tonra J. 2014. “Using distributional semantics to trace influence and imitation in romantic orientalist poetry.” In AHA!-orkshop 2014 on Information Discovery in Text. ACL.

Burgess and Lund 1997 Burgess, Curt, and Kevin Lund. 1997. “Representing abstract words and emotional connotation in a high-dimensional memory space.” Proceedings of the Cognitive Science Society. 1997.

Chaski 2005 Chaski C. E. 2005. “Who’s at the keyboard? Authorship attribution in digital evidence investigations.” *International Journal of Digital Evidence*, 4(1).

De Vel et al. 2001 De Vel O., Anderson A., Corney M., and Mohay G. 2001. “Mining e-mail content for author identification forensics.” *ACM Sigmod Record*, 30(4):55–64.

Deerwester et al. 1990 Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K.; Harshman, Richard. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science*. 41 (6): 391–407

Dell’Orletta et al. 2014 Dell’Orletta F., Venturi G., Cimino A., and Montemagni S. 2014. “T2k2: a system for automatically extracting and organizing knowledge from texts.” In LREC, pages 2062–2070.

Firth 1957 Firth J. R. 1957. “Modes of meaning.” Papers in Linguistics.

Frantzeskou et al. 2006 Frantzeskou G., Stamatatos E., Gritzalis S. and Katsikas S. 2006. “Effective identification of source code authors using byte-level information.” In *Proceedings of the 28th International Conference on Software Engineering* (pp. 893-896).

Gamon et al. 2004 Gamon M. 2004. “Linguistic correlates of style: authorship classification with deep linguistic analysis features.” In *Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics.

Grant 2007 Grant T. D. 2007. “ Quantifying evidence for forensic authorship analysis.” *International Journal of Speech-Language and the Law*, 14(1), 1 -25.

Grieve 2007 Grieve J. 2007. “Quantitative Authorship Attribution: An Evaluation of Techniques.” *Literary and Linguistic Computing*, 22(3):251–270, May.

Halliday 1994 Halliday M. A. K. 1994. *Functional grammar.* London: Edward Arnold.

Harris 1970 Harris Z. S. 1970. *Distributional structure.* Springer.

Herbelot 2015 Herbelot A. 2015. “The semantics of poetry: A distributional reading.” *Digital Scholarship in the Humanities*, 30(4):516–531.

Hirst and Feiguina 2007 Hirst G. and Feiguina O. 2007. “Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts.” *Literary and Linguistic Computing*, 22(4):405–417, September.

Holmes 1994 Holmes D. I. 1994. “ Authorship attribution.” *Computers and the Humanities*, 28, 87–106.

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

Juola 2004 Juola P. 2004. “ Ad-hoc authorship attribution competition.” In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (pp. 175-176).

Koppel and Schler 2004 Koppel M. and Schler J. 2004. “ Authorship verification as a one-class classification problem.” In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM.

Kruszewski and Baroni 2014 Kruszewski G. and Baroni M. 2014. “Dead parrots make bad pets: Exploring modifier effects in noun phrases.” *Lexical and Computational Semantics* (* SEM 2014), page 171.

Landauer 2007 Landauer, Thomas K. 2007. “LSA as a theory of meaning.” *Handbook of latent semantic analysis* 3 (2007): 32.

Landauer and Dumais 1997 Landauer, Thomas K., and Susan T. Dumais 1997. “A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.” *Psychological review* 104.2 (1997): 211.

Lenci 2008 Lenci A. 2008. “Distributional semantics in linguistic and cognitive research.” *Italian journal of linguistics*, 20(1):1–31.

Li et al. 2006 Li J., Zheng R., and Chen H. 2006. “From fingerprint to writeprint.” *Communications of the ACM*, 49(4):76–82.

Miller and Charles 1991 Miller G. A. and Charles W. G.. 1991. “Contextual correlates of semantic similarity.” *Language and cognitive processes*, 6(1):1–28.

Mosteller and Wallace 1964 Mosteller F. and Wallace D. L. 1964. “Inference and disputed authorship: The Federalist.” Addison-Wesley.

Rudman 1997 Rudman J. 1997. “The state of authorship attribution studies: Some problems and solutions.” *Computers and the Humanities*, 31(4):351–365.

Schütze 1992 Schütze H. 1992. “ Dimensions of meaning.” In Supercomputing’92, Proceedings, pages 787–796. IEEE.

Schütze 1993 Schütze H. 1993. “ Word space.” In Advances in Neural Information Processing Systems 5. Citeseer.

Stamatatos 2006 Stamatatos E. 2006. “ Authorship attribution based on feature set subspacing ensembles.” *International Journal on Artificial Intelligence Tools*, 15(05):823–838.

Stamatatos 2009 Stamatatos E. 2009. “A survey of modern authorship attribution methods.” *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, March.

Stamatatos et al. 2000 Stamatatos E., Fakotakis N., and Kokkinakis G. 2000. “Automatic text categorization in terms of genre and author.” *Computational linguistics*, 26(4):471–495.

Stamatatos et al. 2001 Stamatatos E., Fakotakis N., and Kokkinakis G. 2001. “Computer-based authorship attribution without lexical measures.” *Computers and the Humanities*, 35(2):193–214.

Teng et al. 2004 Teng G., Lai M.S., Ma J.B., and Li Y. 2004. “E-mail authorship mining based on SVM for computer forensics.” In Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, volume 2, pages 1204–1207. IEEE.

Uzuner and Katz 2005 Uzuner O. and Katz B. 2005. “A comparative study of language models for book and author recognition.” In Natural Language Processing–IJCNLP 2005, pages 969–980. Springer.

Van Halteren et al. 2005 Van Halteren H., Baayen H., Tweedie F., Haverkort M., and Neijt A. 2005. “New machine learning methods demonstrate the existence of a human stylome.” *Journal of Quantitative Linguistics*, 12(1):65–77.

Yule 1938 Yule G. U. 1938. “On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship.” *Biometrika*, 30, 363-390.

Yule 1944 Yule G. U. 1944. *The statistical study of literary vocabulary.* Cambridge University Press.

Zhao and Zobel 2005 Zhao Y. and Zobel J. 2005. “Effective and scalable authorship attribution using function words.” In *Information Retrieval Technology*, pages 174–189. Springer.

Zheng et al. 2006 Zheng R., Li J., Chen H., and Huang Z. 2006. “A framework for authorship identification of online messages: Writing-style features and classification techniques.” *Journal of the American Society for Information Science and Technology*, 57(3):378–393, February.

Zipf 1932 Zipf G. K. 1932. *Selected studies of the principle of relative frequency in language.* Harvard University Press, Cambridge, MA.