Is Starnone really the author behind Ferrante?

Jacques Savoy
University of Neuchatel, Switzerland

Abstract

Elena Ferrante is a pen name known worldwide, authoring novels such as the bestseller My Brilliant Friend. A recent study indicates that the true author behind these books is probably Domenico Starnone. This study aims to select a set of approved authorship methods and appropriate feature sets to prove, with as much certainty as possible, that this conclusion is correct. To achieve this, a corpus of contemporary Italian novels has been generated, containing 150 books written by forty authors (including seven by Ferrante). Six authorship identification models have been applied to this data set (Delta, Labbé's distance, nearest shrunken centroids (NSC), naïve Bayes, *k*-nearest neighbors, and character *n*-grams). Using either an instance- or profile-based matching technique, the same result (Starnone) appears very often in first place. Modifying the feature set to include between 50 and 2,000 of the most frequent tokens or lemmas does not change this result. When removing Starnone's novels from the corpus, all approved attribution methods tend to indicate different names as the most probable author. This result confirms not only that the outputs of these methods are independent but also that the true author is certainly Starnone. Finally, a lexical analysis reveals the reasons justifying this conclusion.

Correspondence:
Jacques Savoy, University of Neuchatel, rue Emile Argand 11, 2000 Neuchatel, Switzerland.
E-mail:
Jacques.Savoy@unine.ch

1 Introduction

With the translation of the successful novel My Brilliant Friend into many languages, the pen name 'Elena Ferrante' has recently gained worldwide attention. But who is behind this name? In Italy in particular, several names have been proposed: mainly well-known female novelists originating from Naples (e.g. Milone, Parrella, Ramodino), but also some men (De Luca, Piccolo, Prisco, etc.), and even a translator and journalist (Raja). These suggestions have been formulated by columnists or literary scholars with some intuition about stylistic similarities, and, in Raja's case, according to royalties received. On the other hand, there is a recent scientific study investigating Ferrante's style with statistical tools and methods, concluding that Domenico Starnone is the writer whose profile is closest to that of Ferrante (Tuzzi and Cortelazzo, unpublished).

This recent Italian case is not exceptional. The first mention of a pseudonym in literature is the name 'Nobody', employed by Ulysses in the Odysseus. Many other examples can be found, even recently with *The Cuckoo's Calling*, a novel published in 2013 by R. Galbraith, whose real author is J. K. Rowling (Juola, 2016). In France, in the 1970s, literary critics wrote that Romain Gary's style was 'boring' or 'has been'. In 1973, Gary published the novel Gros Câlin under the pen name Emile Ajar, a book hailed as 'fresh and new style' by the press (Labbé, 2008). In other cases, a newly discovered poem should, if possible, be attributed to his/her true author (Thisted and Efron, 1987). Similar disputed cases occur in other fields, such as the attribution of the Book of Mormon to

Digital Scholarship in the Humanities, Vol. 33, No. 4, 2018.

doi:10.1093/llc/fqy016 Advance Access published on 22 May 2018

J. Smith (Jockers, 2013) or the Federalist Papers (Jockers and Witten, 2010; Savoy, 2013). With Shakespeare's works, the common question is to identify passages written by Shakespeare or by a second possible author (Craig and Kinney, 2009).

In the current case, the conclusion reached by Tuzzi and Cortelazzo (unpublished) is supported by a computer-based attribution model and by comparing seven of Ferrante's books with books by thirty-nine other Italian novelists. The current study expands this investigation by considering additional modern attribution methods (Juola, 2008; Stamatatos, 2009; Jockers and Witten, 2010; Juola, 2016) to identify the true writer behind the pen name Ferrante.

Is it possible to estimate the reliability of each of these computed attributions? Is a single method enough to convince a court that the suggested attribution is correct (Chaski, 2013)? In information retrieval (IR) (Voorhees and Harman, 2005), to prove that one IR model is significantly better than another, this conclusion must be supported by more than one test collection. When combining different authorship attribution results (Juola, 2008), can we be practically certain that the proposed decision is correct? The current study focuses on these questions and suggests a methodology verifying that Starnone is the real author of Ferrante's novels.

To achieve this objective, this study assumes that each novel in our corpus was written by a single person. Of course, collaboration might be responsible for the addition of detail to a romance, the development of a character, or the inclusion of a dialog, but the writing itself is the brainchild of a single person. Second, the name associated with a book is the true writer, and therefore all novels published under this name were written by that person. Finally, our attribution schemes will only take account of textual elements. No other metadata information (the author must be a female, must have lived in Naples, etc.) will be considered when proposing attributions.

Finally, to address this question, it should be recognized that writing style is not fully determined by the author and his/her background (gender, age, social origin, nationality, psychological traits, etc.). Such writeprints are also influenced by the text's genre (the style adopted in an essay is different from that of a poem), period (we do not write nowadays as we did in the 70s), topic (determining some of the vocabulary chosen), type (oral, written, or Web-based), and audience (formal or colloquial).

The rest of this article is organized as follows. The next section describes the Italian corpus used to discover Ferrante's true identity. Section 3 presents the six attribution models applied in our study, state of the-art among authorship attribution methods. Section 4 applies these to the Italian collection and reveals the author of Ferrante's books. Section 5 presents a new verification protocol to ensure that our conclusion (that Starnone is the true writer of Ferrante's novels) is correct. Section 6 gives a more detailed lexical analysis, explaining the close similarity found by the selected attribution methods. A conclusion presents the main findings of this study.

2 Corpus of Italian Novels

Our investigation to discover the real writer behind Ferrante is based on a corpus of contemporary Italian novels called PIC (Padua Italian Corpus). This collection was generated by a team of researchers at the University of Padua, under the supervision of Prof. Michelle Cortelazzo and Prof. Arjuna Tuzzi (Tuzzi and Cortelazzo, unpublished). Table 1 presents the list of authors appearing in this collection, together with their gender and the number of novels included in the corpus. As shown below, the PIC contains 150 books dedicated to adult readers, to ensure a similar genre of text. Forty different authors (twenty-seven men, twelve women, and one Ferrante) are included, each appearing with at least two works, and as many as ten in Starnone's case. Ferrante is included, with seven books (including the four novels of her tetralogy My Brilliant Friend). A careful editing process has been undertaken to remove all elements not belonging to the text itself (page numbers, running titles, etc.), as well as a thorough checking of the spelling.

During the selection of this corpus, all the names suspected to be behind the pseudonym Ferrante

Table 1 Author name, gender (M/F), and the number of novels

Name	Gender	Number	Name	Gender	Number
Affinati	M	2	Montesano	M	4
Ammaniti	M	4	Morazzoni	F	2
Bajani	M	3	Murgia	F	5
Balzano	M	2	Nesi	M	3
Baricco	M	4	Nori	M	3
Benni	M	3	Parrella	F	2
Brizzi	M	3	Piccolo	M	7
Carofiglio	M	9	Pincio	M	3
Covacich	M	2	Prisco	M	2
De Luca	M	4	Raimo	M	2
De Silva	M	5	Ramondino	F	2
Faletti	M	5	Rea	M	3
Ferrante	?	7	Scarpa	M	4
Fois	M	3	Sereni	F	6
Giordano	M	3	Starnone	M	10
Lagioia	M	3	Tamaro	F	5
Maraini	F	5	Valerio	F	3
Mazzantini	F	4	Vasta	M	2
Mazzucco	F	5	Veronesi	M	4
Milone	F	2	Vinci	F	2

have been included. Furthermore, an effort has been made to include more women novelists. In addition, many books have been written by ten authors from Naples and the Campania region (namely, De Luca, De Silva, Milone, Montesano, Parrella, Piccollo, Prisco, Ramondino, Rea, and Starnone). This regional element is important in Italian, due to the presence of spelling differences between regions (diatopic variation), and the use of dialect-specific words and expressions.

In total, the corpus contains 9,609,234 word-tokens with an average of 64,062 tokens/novel (standard deviation: 38,228). The largest book is composed of 196,914 tokens (Faletti, Io uccito, 2002), and the smallest of 7,694 tokens (written by Parrella, Behave, 2011, and the only work with fewer than 10,000 word-tokens). For Ferrante's novels, the average size is 88,933 word-tokens (min: 36,222 (La figlia oscura), max: 138,622 (Storia della bambina perduta)). In total, Ferrante's writings represent 6.48% of the corpus, while those of Faletti constitute the largest share (6.6%) followed by Starnone (6.4%), and Mazzucco (6.15%). The smallest contribution is provided by Parrella (0.36%), followed by Vinci (0.58%) and Nori (0.64%).

As preprocessing, and for all experiments, each novel's text has been analyzed by the TreeTagger POS tagger[1] to derive both the word-tokens (tokenization) and the lemmas (dictionary entries). When the lemma cannot be defined by the tagger, the corresponding token is used (usually when dealing with names). Then all uppercase letters are transformed to their lowercase equivalents, and all punctuation marks and digits are removed. This decision is justified not only by the fact that punctuation marks can take different graphical forms (e.g. ‘‘, ’’, ", «, and ») but also that they can be imposed or modified by the editor or publisher.

From a computer-based attribution perspective, it is worth mentioning that this corpus has three essential characteristics. First, each text contains more than 10,000 word-tokens (with a single exception); second, a rigorous spell-checking process has been applied; third, according to best practice and established protocols (Juola, 2016), additional obfuscating factors have been isolated. Therefore, the corpus contains works of the same language (Italian), genre of text (novels for adult readers), and approximate period (from 1987 to 2016). The selection of the authors to be included takes account of the region, and both genders are represented by many books.

3 Authorship Attribution Methods

To determine the true author of a text, numerous authorship attribution methods have been proposed (e.g. Juola and Vescovi (2011) suggest more than 1,000 approaches). Therefore, it may be hard to believe that a single attribution model could always provide the correct answer in all circumstances. According to the no free lunch theorem (Wolpert, 2001), averaged over all possible problems, every classification algorithm has a similar accuracy rate when classifying new unseen data. Thus, no learning scheme can be universally better than all the others. Therefore, to ascertain a proposition with a higher degree of certainty, several approaches must be taken into account. Such an evaluation methodology has been suggested by Juola (2016). To be accepted in a US court (Chaski, 2013), such

methods must reflect the state of the art in the domain. They must have demonstrated their effectiveness and robustness in several contexts using different test collections.

To respect these constraints, the well-known Delta authorship attribution model has been selected. Proposed by Burrows (2002), several variants have been suggested and evaluated (Hoover, 2004a, 2004b; Burrows, 2007), and the model's theoretical properties have been analyzed (Argamon, 2008). It has been used in various attribution evaluation studies Jockers et al. (2008), Savoy (2012), Eder (2015), Savoy (2016), Evert et al. (2017), Kocher and Savoy (unpublished), and Jockers and Witten (2010) showed that the Delta method could surpass the performance levels achieved by the SVM method in authorship attribution.

This attribution model makes use of a subsection of the vocabulary and considers only the most frequent word-types (MFT) or lemmas (MFL) (containing mainly function words such as determiners, pronouns, prepositions, conjunctions, and some auxiliary verb forms). The number of terms (word-tokens or lemmas) to be included is not precisely defined, but the norm is to consider a value between 200 and 500 terms, determined without making reference to Ferrante's novels (Savoy, 2015). To weight each selected term t_i (denoted Z score(t_{ij})), its relative term frequency rtf_{ij} in a text T_j is computed alongside the mean ($\overline{rtf_i}$) and standard deviation (s_i) of that term over all novels in the corpus (see Equation (1)).

$$\text{Zscore}(t_{ij}) = {(rtf_{ij} - \overline{rtf_i})}/{s_i}. \quad (1)$$

Then, given a disputed text Q, an author profile A_k (concatenation of all his/her writings), and a set of terms t_i, for $i = 1, 2, \ldots, m$, the Delta distance value (denoted $\Delta(Q, A_k)$) is computed according to Equation (2).

$$\Delta(Q, A_k) = {1}/{m} \cdot \sum_{i=1}^{m} \left| Zscore(t_{iQ}) - Zscore(t_{ik}) \right|. \quad (2)$$

When for one term both Z scores are large and have opposite polarity, the distance will grow. In this case, one author tends to use the corresponding term more frequently than the mean, while the other employs it rarely. When for all terms the Z score values are very similar, the distance between the two texts will be reduced. The smallest distance over all authors A_k (for $k = 1, 2, \ldots, r$) determines the proposed true author.

Second, the intertextual distance measure suggested by Labbé (2007) was chosen. The effectiveness of this model has been evaluated by different studies reflecting different contexts (Labbé, 2008; Tuzzi, 2010; Savoy, 2012; Ratinaud and Marchand, 2016; Kocher and Savoy, 2017; Tuzzi and Cortelazzo, unpublished). For example, in Labbé and Labbé (2013), this distance function is applied to detect duplicate and automatically generated scientific articles. This intertextual distance returns a value between 0.0 and 1.0 depending on the lexical overlap between two texts. When two texts are identical, the distance is 0.0. The largest distance of 1.0 would appear when the two books have nothing in common (e.g. one in Italian and the other in Chinese). Between these two limits, the distance value depends on the number of terms appearing in both novels and their occurrence frequencies.

More precisely, the distance between Text A and Text B (denoted $D(A, B)$) is computed according to Equation (3), where n_A indicates the length of Text A (in number of tokens), and tf_{iA} denotes the absolute frequency of the ith term (for $i = 1, 2, \ldots, m$). The value m represents the vocabulary length. It is rare that both texts have the same length, so let us assume that Text B is the longer. To reduce the longer text to the size of the smaller, each of the term frequencies (in our case tf_{iB}) is multiplied by the ratio of the two text lengths, as indicated in the second part of Equation (3).

$$\text{D(A, B)} = \sum_{i=1}^{m} \left| tf_{iA} - \widehat{tf_{iB}} \right| \Big/ 2\, n_A \quad \text{with} \quad \widehat{tf_{iB}} = tf_{iB} \cdot {n_A}/{n_B}. \quad (3)$$

The third model chosen is the nearest shrunken centroids (NSC) method (Tibshirani et al., 2002; Tibshirani et al., 2003), judged to be an effective approach in authorship attribution (Jockers and

Witten, 2010; Jockers, 2013; Kocher and Savoy, (unpublished)). This strategy can be viewed as a variant of the *k*-nearest neighbors (*k*-NN) method, in which less discriminative features are ignored (small feature weights are shrunken toward 0, an idea drawn from the ridge and lasso regression model). Without providing all the details (see Tibshirani et al., 2003), the general idea is as follows. For each term t_i (for $i=1, 2, \ldots, m$) appearing in a novel $T_j = 1, 2, \ldots n$, its relative term frequency rtf_{ij} is estimated. From these values, its mean across all novels in the corpus (denoted $\overline{rtf_i}$) is computed, as well as its mean across all novels written by author A_k (for $k = 1, 2, \ldots r$) (indicated by $\overline{rtf_{ik}}$). Where n_k is the number of novels written by A_k, the standard deviation (denoted s_i) is also determined. From these values, a discriminative value (denoted w_{ik}) for term t_i and author A_k is calculated according to Equation (4).

$$w_{ik} = \left(\overline{rtf_{ik}} - \overline{rtf_i}\right) \Big/ (m_k \cdot s_i) \text{ with } m_k = \sqrt{1/n_k + 1/n}. \quad (4)$$

The principle behind this feature weighting scheme can be explained as follows. When the mean of the *i*th term for the *k*th author ($\overline{rtf_{ik}}$) is similar to the overall mean for this term ($\overline{rtf_i}$), the resulting feature weight for this author is small. This feature does not have the discriminative power to distinguish the underlying writer from all the others. The amplitude of this difference must, however, be analyzed according to frequency distribution and is thus divided by its standard deviation ($m_k \,.\, s_i$). The final weight w'_{ij} associated with the *i*th term in the *j*th novel is shown in Equation (5), in which Δ (the shrinkage parameter) is a constant, and $(v)+$ a function returning the value v if $v > 0$, and otherwise 0.

$$w'_{ij} = \text{sign}(w_{ij}) \cdot (|w_{ij}| - \Delta)+. \quad (5)$$

According to this formula, all term weights are decreased by the same Δ value. Moreover, a feature weight should be set to 0 when it is smaller than Δ. Such features are viewed as generating more noise, rather than as being helpful. Increasing the value of Δ will reduce the number of features taken into account. Usually, a given feature is useful when discriminating between a few possible authors (or categories), rather than all. Thus, w'_{ij} values for a few authors are different from 0. Moreover, when for the *i*th term, all the corresponding weights w'_{ij} are set to 0, this feature is ignored for all attributions. This happens when the overall mean ($\overline{rtf_i}$) for that term is very similar across all authors ($\overline{rtf_{ik}}$).

Fourth, as a typical text classifier derived from the machine learning paradigm, the naïve Bayes model (Mitchell, 1997) has been chosen. This method was the first proposed to solve authorship attribution (Mosteller and Wallace, 1964), and it is usually suggested as an effective baseline for evaluating machine learning algorithms (Witten et al., 2016). Of course, other studies have proposed applying the naïve Bayes as an authorship identification method, such as Juola and Vescovi (2011) and Savoy (2012), in which this method proves highly effective.

Having a set of possible authors (or hypotheses) denoted by A_k for $k = 1, 2, \ldots r$, the naïve Bayes model combines the prior probability that a given author wrote the disputed text (denoted by $\text{Prob}[A_k]$) and the likelihood probability. The latter is defined as the product of observing all terms t_i (for $i = 1, 2, \ldots, m$), knowing that the text is written by the author A_k. This formulation assumes that the term distributions are independent, which is unrealistic or naïve. For a query text Q, the naïve Bayes model selects as the probable author the writer who maximizes Equation (6), in which t_i represents the *i*th term included in the query text Q, and n_Q indicates the size of the query text.

$$Arg\,max_{A_k} \text{Prob}[A_k|Q] = \text{Prob}[A_k] \cdot \prod_{i=1}^{n_Q} \text{Prob}[t_i|A_k]. \quad (6)$$

To estimate the prior probabilities ($\text{Prob}[A_k]$), one can choose either a uniform distribution over all possible authors, or a distribution according to the proportion of novels written by each author. To determine the term probability, all texts belonging to the same author are concatenated to generate the corresponding profile. For each term t_i, this probability is estimated as the ratio between its

occurrence frequency in the profile A_k (tf_{ik}) and the size of this sample (n_k), as shown in Equation (7).

$$\text{Prob}[t_i \mid A_k] = {}^{\text{tf}_{ik}}\!/_{n_k}. \quad (7)$$

This definition tends, however, to over-estimate the probabilities of terms occurring in the text with respect to missing terms. For the latter, the occurrence frequency (and probability) was 0, so a smoothing approach had to be applied to correct this. As for the other methods, we will apply Lidstone's smoothing method, which estimates $\text{Prob}[t_i|A_k] = (tf_{ik}+\lambda)/(n_k + \lambda|V|)$, with λ as a parameter (set to 0.1 in this study), and $|V|$ indicating the vocabulary size.

In addition, the *k*-NN approach, taken from the domain of machine learning, represents a non-parametric model in which each text is viewed as a point in an *m*-dimensional space (instance-based model). Each of these dimensions corresponds to a feature or a token (lemma) in the current study. The relative frequency of each term indicates the amplitude in the corresponding direction. The same representation is applied to the disputed document.

To measure the distance between two points, numerous functions have been suggested, based on the L^1 norm (e.g. Manhattan, see Equation (8)), L^2 norm (e.g. Euclidian distance), inner product (e.g. Dice), entropy-based (e.g. Kullback–Leibler divergence), or on *ad hoc* principles (combining two or more measures). Such a strategy was proposed for authorship attribution (AA) (Zhao and Zobel, 2005; Savoy, 2012).

$$D_{Manhattan}(A, Q) = \sum_{i=1}^{m} \left| rtf_{iA} - rtf_{iQ} \right|. \quad (8)$$

In a recent study using this attribution method, Kocher and Savoy (2017) found that both the Tanimoto (see Equation (9)) and Matusita function (Equation (10)) are effective in profiling the author of a text. In these formulations, the text with known authorship (or the author profile) is denoted by *A*, while the disputed text is represented by *Q*.

$$D_{Tanimoto}(A, Q) = \sum_{i=1}^{m} \left| rtf_{iA} - rtf_{iQ} \right| \Big/ \sum_{i=1}^{m} Max\left(rtf_{iA}, rtf_{iQ}\right) \quad (9)$$

$$D_{Matusita}(A, Q) = \sqrt{\sum_{i=1}^{m} \left(\sqrt{rtf_{iA}} - \sqrt{rtf_{iQ}}\right)^2}. \quad (10)$$

As the last attribution model, the letter *n*-gram (Abbasi and Chen, 2008) has been used to represent the different author profiles A_k (concatenation of all their novels). This text representation is more difficult to interpret for the user but tends to be highly effective (Juola, 2008; Koppel et al., 2009), as demonstrated by the last PAN CLEF evaluation campaigns (Potthast et al., 2017).

When generating this text surrogate, overlapping token *n*-grams (for $n = 1, 2, \ldots, 5$) are applied. A word boundary does not prevent *n*-gram generation, and each boundary appears in the resulting *n*-gram as a space. For example, the phrase 'il mia amica è' (my friend is) produces the following 3-grams {'_il', 'il_', '_mi', 'mia', 'ia_', . . ., 'ca_', '_è_'}, where the spaces are underlined. However, sentence boundaries are respected and block *n*-gram generation. Given that not all texts have the same length, the representation is usually based not on the absolute frequency (*tf*) but the relative frequency (*rtf*). Finally, the distance computation between the representation of the disputed text and the author profile can be established according to the Manhanttan, Tanimoto, or Matusita functions.

4 Evaluation

Having six approved attribution models to identify the true author, the next step is to apply them with an appropriate feature set. To achieve this objective, the seven novels written by Elena Ferrante form the test set, and all other 143 books belong to the training set. Using the Delta model, all books written by an author are concatenated to build the corresponding author's profile. Applying this approach, all seven of Ferrante's novels are assigned to Domenico Starnone, taking account of the 50,

Table 2 Ranked lists produced by the Delta model (100 MFT, profile-based approach)

Rank	Distance	Author	Distance	Author	Distance	Author
	L'amore molesto		**L'amica geniale**		**Storia bambina perduta**	
1	0.602	Starnone	0.528	Starnone	0.565	Starnone
2	0.731	Brizzi	0.629	Balzano	0.721	Veronesi
3	0.785	Tamaro	0.676	Sereni	0.745	Balzano
4	0.825	Sereni	0.680	Veronesi	0.747	Sereni
5	0.828	Milone	0.730	Carofiglio	0.768	Carofiglio

Table 3 Ranked lists produced by the Delta model (200 MFT, profile-based approach)

Rank	Distance	Author	Distance	Author	Distance	Author
	L'amore molesto		**L'amica geniale**		**Storia bambina perduta**	
1	0.650	Starnone	0.524	Starnone	0.539	Starnone
2	0.806	Brizzi	0.686	Veronesi	0.709	Veronesi
3	0.837	Milone	0.700	Balzano	0.717	Carofiglio
4	0.850	Tamaro	0.721	Brizzi	0.783	Balzano
5	0.874	Lagiola	0.726	Milone	0.792	Tamaro

100, 150, 200, 250, 300, 400, 500, 1,000, 1,500, or 2,000 MFT or MFL. As the Italian language has a richer morphology than English, the lemma can reduce some redundant variations present in the tokens (e.g. from the tokens amico, amica, amici, the same lemma (amico, friend) is derived).

Usually, values between 200 and 500 MFT form a feature set found to be effective in determining the real author of a text or an excerpt of a novel (Savoy, 2015). Those sizes also correspond to values indicated in the seminal paper on this approach (Burrows, 2002) or (Hoover, 2004a). When it comes to authors whose style might be seen as similar to Ferrante's, one can find Veronesi (40 times), Milone (27), Carofiglio (25), Brizzi (24), Balzano (10), Tamaro (10), Mazzucco (7), Sereni (7), Giordano (2), Lagiola (1), or Parrella (1) ranked second or third, according to the 154 tests (7 novels $\times$ 11 feature sets $\times$ 2 ranks).

Table 2 reports the top five names, sorted by the Delta model using the 100 MFT, with three of Ferrante's novels, namely, L'amore molesto (her first novel, published in 1992), L'amica geniale (the first book of her tetralogy, 2011), and Storia della bambina perduta (her most recent book, 2014). Table 3 depicts the same information but obtained with the 200 MFT.

In both cases, it is interesting to note that the distance value difference between the first and the second author is larger than the difference between the second and third. For example, in Table 2, the difference between the first two ranks is 0.731–0.602 = 0.129 (or 21.4%). The divergence between the second and the third is 0.785–0.731 = 0.054 (or 7.4%). This comparison indicates that the first answer is clearly more probable than the rest of the ranked list. A similar finding can be found in Table 3. When considering the eleven MFT sets, the average difference between the first two ranks is 31.3% compared to 3.5% between the second and third ranks.

With the distance-based attribution model (Labbé, 2007), the entire vocabulary generates the feature set, using either word-tokens or lemmas (but without punctuation marks). When building each novel's representation, the terms with an occurrence frequency of one or two are removed. Then, the intertextual distance is computed for all pairs of novels and ranked from the smallest to the largest distance. Table 4 gives an excerpt of such a ranked list.

Table 4 Ranked list of Labbé's distances between two novels (tokens)

Rank	Distance	Author	Title	Author	Title
s	0.126	Ferrante	Storia di chi fugge e di chi resta	Ferrante	Storia della bambina perduta
2	0.130	Ferrante	Storia del nuovo cognome	Ferrante	Storia di chi fugge e di chi resta
3	0.136	Ferrante	L'amica geniale	Ferrante	Storia del nuovo cognome
4	0.138	Ferrante	Storia del nuovo cognome	Ferrante	Storia della bambina perduta
5	0.145	Veronesi	Caos calmo	Veronesi	Terre rare
...	...	...	...	...	...
40	0.212	Ferrante	I giorni dell abbandono	Starnone	Prima esecuzione
41	0.213	Carofiglio	Le perfezioni provvisorie	Carofiglio	Bordo vertiginoso delle cose
42	0.213	Ferrante	L'amore molesto	Starnone	Eccesso di zelo
...	...	...	...	...	...
48	0.218	Ferrante	La figlia oscura	Starnone	Prima esecuzione
...	...	...	...	...	...
54	0.221	Ferrante	Storia di chi fugge e di chi resta	Starnone	Prima esecuzione
...	...	...	...	...	...
65	0.228	De Silva	Mia suocera beve	Veronesi	Caos calmo
66	0.228	Ferrante	Storia di chi fugge e di chi resta	Starnone	Lacci
...	...	...	...	...	...
75	0.231	Ferrante	Storia del nuovo cognome	Milone	Il silenzio del lottatore
...	...	...	...	...	...
96	0.237	Raimo	Latte	Starnone	Prima esecuzione

In the top ranks, and with smallest distances, are novels written by the same author, as shown in Table 4. The top four links correspond to Ferrante's novels, published in the space of 1 year, and dealing with similar topics (My Brillant Friend tetralogy). As the distance value increases, the certainty that both books are written by the same novelist decreases. Ranked at #40 is the first 'incorrect' pairing (with a distance of 0.212); the second such pairing, between a novel written by Ferrante and one by Starnone, is ranked at #42 (distance 0.213). Subsequently, there are two additional 'incorrect' pairings (Rank #48 and #54, both pairing a novel by Ferrante and a novel by Starnone), before the first genuinely erroneous link (between De Silva and Veronesi), ranked at #65. This result indicates that the real author behind Elena Ferrante's writings is, with some certainty, Domenico Starnone. Having four additional 'incorrect' pairings between these two names before any erroneous link between two other authors is clearly surprising.

Taking account of lemmas instead of tokens, a similar ranked list can be established, with its first pairing between a Ferrante novel (Storia della bambina perduta) and a Starnone novel (Lacci) ranked #35 (with a distance of 0.181).

The next step is to estimate the probability that such a small distance value (0.212 in this study) could be observed between two texts written by the same person. To establish this probability, one can model the distance values (some examples are given in Table 4) as derived from a mixture of two Gamma distributions, one between two novels written by the same person (distribution D1 in Fig. 1), the second with pairs linking papers produced by two distinct authors (distribution *D*2). The estimation of both distributions has been completed without Ferrante's novels. The Gamma distribution is chosen because the distance values are never negative, can take all positive values, and are skewed to the right. In Savoy (2016), the author suggests two Beta distributions with possible values limited within the range [0, 1].

To estimate the probability that a distance value $d = 0.212$ links two texts written by the same author, Prob $[0.212-\delta < d < 0.212+\delta \mid D1]$ and Prob $[0.212-\delta < d < 0.212+\delta \mid D2]$ (with δ fixed to 0.05) is computed. The sum of these two probabilities is used as a normalization constant. With this method, the probability that $d = 0.212$ links two texts written by the same person is given by Prob

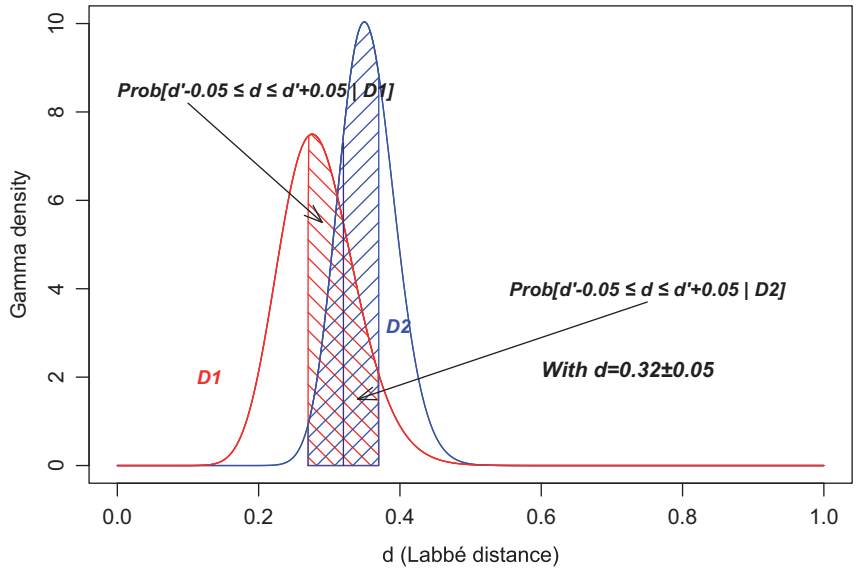

Fig. 1 Labbé's distance viewed as two Gamma distributions based on distance values (token text representation)

$[0.212-\delta < d < 0.212+\delta \mid D1]$ and is equal to 0.96 in our case.

Instead of considering the entire ranked list, one can just look at the closest books for each of Ferrante's novels (ignoring all Ferrante's other books). In all cases, the first rank is occupied by a book written by Starnone. In the second position (ignoring Starnone's works), one can usually observe a pairing with a book written by Milone (Il silenzio del lottatore), and, for the first of Ferrante's novels, a book by De Luca (Tu, mio). With lemmas, the same book by Milone occurs 4 times, while for the first and last novel of the My Brilliant Friend tetralogy, a novel authored by Sereni (Una storia chiusa) appears. With the third book of this tetralogy, the closest novel is Caos calmo written by Veronesi.

As a third attribution method, the NSC model has been applied. As feature sets, the 100, 150, 200, 250, 300, 400, 500, 1,000, 1,500, and 2,000 MFT or MFL are considered. In all cases, the seven novels by Ferrante have been assigned to Starnone when the shrunken parameter Δ is fixed to 0.2, 0.5, or 0.7. With 500 MFT and Δ set to 0.5, decisions are reached with 23.4% positive feature weights, 29.3% negative, and 47.3% (or 236 of 500) set to 0. With 300 MFT, the distribution of the feature weights is similar (25.4% positive, 31.6% negative, and 42.9% removed). Table 5 reports the top five names, sorted by the NSC model using the 200 MFT, for three Ferrante novels (L'amore molesto, L'amica geniale, and Storia della bambina perduta). As already shown in Tables 2 and 3, the difference between the distance in the first and second rank is higher than that between the second and third rank.

With a larger Δ value than 0.7, the feature set size is significantly reduced. Compared to our previous example, with 300 MFT and the shrinkage parameter Δ fixed at 2.0, the decisions are then reached with 4.7% positive weights, 1.9 % negative, and with 93.4% are set to 0. As a direct effect, attributions to authors other than Stanone appear. For example, using 300 MFT or MFL ($\Delta = 2.0$), the first two novels written by Ferrante (L'amore molesto and I giorni dell abbandono) are assigned to Tamaro. The NSC approach can, however, provide an estimation that the proposed solution is the correct one (Tibshirani et al., 2003). In our last example (300 MFT, $\Delta = 2.0$), the probability estimated for assigning L'amore molesto to Tamaro is 0.37. With conservative parameter settings (e.g. 500 MFT or MFT, $\Delta = 0.5$), the assignment of Ferrante's

Table 5 Ranked lists produced by the NSC model (200 MFT, $\Delta = 0.5$)

Rank	Distance	Author	Distance	Author	Distance	Author
	L'amore molesto		**L'amica geniale**		**Storia bambina perduta**	
1	82.01	Starnone	60.26	Starnone	64.91	Starnone
2	103.88	Tamaro	67.01	Balzano	80.02	Balzano
3	104.52	Milone	75.36	Veronesi	83.93	Veronesi
4	107.64	Balzano	79.85	Milone	92.91	Milone
5	111.65	Brizzi	82.65	Giordano	94.29	Giordano

Table 6 Ranked lists produced by the naïve Bayes model (50 MFT, profile-based approach)

Rank	Author L'amore molesto	Author L'amica geniale	Author Storia bambina perduta
1	Milone	Milone	Starnone
2	Starnone	Starnone	Milone
3	Brizzi	Carofiglio	Carofiglio
4	Carofiglio	Balzano	Balzano
5	Tamaro	Brizzi	Parrella

seven books to Starnone is given a probability between 0.97 and 0.99 (and sometimes even 1.0). This last value is certainly over-estimated but clearly differs from the estimation obtained previously with Tamoro (0.37).

Reducing the number of word-types to 50 ($\Delta = 0.2$ or 0.5), only one assignment differs (although we repeat that considering such a reduced set size is usually not effective, particularly when working with entire novels). In this case, the novel L'amica geniale is assigned to Balzano, while the other six are still attributed to Starnone. With fifty lemmas, this single distinct result does not appear.

With the naïve Bayes model, using a uniform prior distribution over the thirty-nine authors, the feature set corresponds to the 50, 100, 150, 200, 250 300, 400, 500, 1,000, 1,500, and 2,000 MFT. With the exception of the fifty MFT, all other experiments indicate Starnone in first place. With the fifty MFT (see Table 6), one can detect two differences. For the first novel (L'amore molesto), and the first of her tetralogy (L'amica geniale), the naïve Bayes model indicates that the probable author is Milone. Of course, such a reduced feature set size is usually not very effective, particularly when working with entire novels (Savoy, 2015).

Using a prior distribution based on the proportion of books written by each author does not modify these results. Replacing the tokens by the lemmas produces the same overall evaluation. Finally, substituting the author-profile representation with an instance-based one (each novel corresponds to a possible hypothesis), the same conclusion is reached.

With the *k*-NN approach, each novel (according to an instance-based approach) can be represented by different feature sets. In the current experiment, eleven feature sets have been applied (50, 100, 150, 200, 250 300, 400, 500, 1,000, 1,500, and 2,000 MFT). As a distance measure, the Manhattan, Tanimoto, and Matusita functions have been applied. In total, we obtained 231 experiments (11 feature sets $\times$ 3 distance functions $\times$ 7 books). Varying the value of the parameter $k = 1$ or 3, the same conclusion always appears: Starnone is the true author of the novels published under the pen name Ferrante. As indicated in Table 1, eleven novelists appear with only two books, implying that the largest value for k is 3.

Table 7 depicts the top five novels found to be the most similar to three of Ferrante's books (L'amore molesto, L'amica geniale, and Storia della bambina perduta). To compute the distance, the Manhattan function was applied to the 200 MFT. To take the final decision, the value of the parameter k must be fixed, but according to the data shown in Table 7, the same result is returned for $k = 1$, or 3 (and even 5). These data indicate that various novels written by Starnone are similar to the three books by Ferrante. For these three cases, the closest is the same novel (Eccesso di zelo) written in 1993. In the top five most similar, one can also find a book written by Milone (Il silenzio del lottatore).

Table 7 Ranked lists produced by the *k*-NN model (200 MFT, Manhattan, instance-based)

Rank	Distance	Author	Distance	Author	Distance	Author
	L'amore molesto		**L'amica geniale**		**Storia bambina perduta**	
1	0.186	Starnone	0.192	Starnone	0.209	Starnone
	Eccesso di zelo		Eccesso di zelo		Eccesso di zelo	
2	0.197	Starnone	0.201	Starnone	0.209	Starnone
	Denti		Via Gemito		Scherzetto	
3	0.209	Starnone	0.212	Starnone	0.215	Starnone
	Via Gemito		Denti		Lacci	
4	0.250	Starnone	0.215	Milone	0.222	Starnone
	Prima esecuzione		Il silenzio del lottatore		Autobiografia	
5	0.251	Milone	0.227	Starnone	0.222	Starnone
	Il silenzio del lottatore		Autobiografia		Via Gemito	

Applying a character *n*-gram approach, the first pertinent feature set is a combination of character unigrams and bigrams extracted from the tokens. The single-letter frequencies are usually not very informative, and letter bigrams offer a better stylistic representation. Combining both unigrams and bigrams tends to produce more discriminate text surrogates.

In our first set of experiments, Ferrante's profile is used as the query text and compared to the thirty-nine other author profiles, producing a ranked list as output. With this procedure, the three distance functions (Labbé, Tanimoto, or Matusita) always return Starnone as the closest profile. Using a 3-gram to 6-gram text representation, Starnone always appears in first place. Ranked second and third are Carofiglio (6 times), Sereni (6), Mazzucco (4), Tamaro (4) and Veroni (4), according to the 24 tests (4 *n*-grams × 3 distances × 2 ranks).

In a second set of experiments, each of Ferrante's novels forms the query text to be compared with the other thirty-nine authors' profiles (*k*-NN with $k = 1$, profile-based approach). An example of three ranked lists is depicted in Table 8, using the Tanimoto distance with unigram and bigram text representation. In this case, the same name appears in first place.

Applying the two other distance functions, the same result always appears in first place. For each of Ferrante's books, the author appearing in first place is Starnone, with *n*-gram = 1 and 2, 3, 4, or 5. The authors found most frequently in second and third place are Tamaro (54 times), Carofiglio (33 times), Milone (26), Sereni (21), Veronesi (14) and Mazzucco (10) according to the 168 tests (4 *n*-grams × 3 distances × 7 novels × 2 ranks). When text representations are generated with 6-grams, Ferrante's first novel (L'amore molesto) and third novel (and smallest, La figlia oscura) are assigned to Mazzucco (ranked first). For the five others, Starnone appears in first place (as well as in second place for the two novels attributed to Mazzucco). Even if *n*-gram text representation has been found effective, the processing time is greater than for word-based representation. The processing time increases exponentially with the value of *n*. For example, where $n = 6$, the attribution of the seven novels by Ferrante took a mean of 17.8 h, while the same task required 2.4 h with the 5-gram model (54.6 min for 4-grams, and 48 min for 3-grams) or 44 s with the Delta model (or 22 s with the NSC approach).

5 Toward a Rigorous Verification Protocol (Open-Set Assumption)

In the previous experiments, the name Domenico Starnone appears very often as the first possible author in the ranked lists generated by different authorship attribution methods. Having the support of more than one attribution model doubtless reinforces the certainty that the true author behind Elena Ferrante is Starnone. To promote a rigorous

Table 8 Ranked lists produced by the combined letter uni- and bigram representation (Tanimoto distance, profile-based approach)

Rank	Distance	Author	Distance	Author	Distance	Author
	L'amore molesto		**L'amica geniale**		**Storia bambina perduta**	
1	0.066	Starnone	0.060	Starnone	0.057	Starnone
2	0.074	Mazzucco	0.071	Milone	0.074	Carofiglio
3	0.075	Milone	0.072	Mazzucco	0.075	Mazzucco
4	0.077	Prisco	0.075	Montesano	0.078	Scarpa
5	0.081	Lagioia	0.075	Balzano	0.078	Nesi

evaluation protocol, a set of criteria must be clearly met.

First, all selected attribution models must be standard approved methods in the field and have been the subject of several distinct evaluations based on various corpora. Each selected attribution method must have been used with success in various contexts, and its behavior must have been analyzed by different previous studies. A new method can be favored by hidden and unknown characteristics of the underlying test collection.

Second, the test collection must respect certain constraints to deliver consistent results. The Italian corpus (Tuzzi and Cortelazzo, unpublished) used in our experiments is composed of books belonging to the same text genre (novels for adult readers). They contain more than 10,000 words (with one single exception), a size found to be appropriate for the achievement of reliable and stable results (Eder, 2015). The text quality has been checked (e.g. spelling), and additional elements (e.g. page numbers, running titles) have been removed. Finally, all the novels were published in the same time period (1987–2016).

Based on these two fundamental elements, each attribution method works in an ideal situation (Chaski, 2013). According to past evaluations, we can suppose that each model has an accuracy rate of 0.8 (or 80%), a rather conservative value. Consequently, the chance of providing an incorrect assignment is 0.2 (or 20%). Assuming that their results are independent, the chance that two attributions are incorrect is $(1-0.8) \times (1-0.8) = (1-0.8)^2 = 0.04$ (or 4%) (Juola, 2016). With six models, this probability decreases to $(1-0.8)^6 = 0.000064$, fewer than 1 in 10,000. Assuming that two of the proposed feature sets are independent (six methods with two feature sets), the probability of a systematic error decreases to $(1-0.8)^{12} \approx 4.1/1$ billion. The chance that such a systematic error might occur is very low.

Until now, the closed-set assumption has been applied: the true author is one of the novelists present in our corpus. To confirm our conclusion, a final step must be included in the evaluation protocol. In this additional stage, all novels written by Starnone are removed from the corpus. In such a case, the real author could be absent from our candidate list. If our assertion is correct (Starnone is the true author), the evaluation situation changes from the closed-set to the open-set context. The true author could be one of the thirty-eight remaining authors, or another author.

After this reduction, the corpus contains thirty-nine authors (with Ferrante) and 140 books. Then, all the previous methods with their feature sets are reapplied. If Starnone is not the true author, another name must appear, more or less recurrently, as the real secret hand behind Elena Ferrante. On the other hand, if several different names appear in the top ranks (limited to first place in this study), our conclusion will be confirmed. In addition, returning several distinct names provides informal proof that the underlying methods are independent.

In this second set of experiments, the Delta model was applied to the seven novels with eleven feature sets with both tokens and lemmas, resulting in 154 tests ($7 \times 11 \times 2$). Using Labbé's distance, seven novels are analyzed with two representations (lemmas and tokens) giving fourteen. The NSC approach gives 924 (7 novels $\times$ 11 features $\times$ 6 distinct Δ values $\times$ 2 tokens and lemmas). The naïve

Table 9 Names ranked first with different methods and parameter settings (corpus without Starnone)

Method	Parameter	Type	Ranked first			
Delta	200 MFT	Profile	Veronesi (3)	Tamaro (2)	Brizzi (1)	Carofiglio (1)
Delta	200 MFL	Profile	Veronesi (4)	Brizzi (2)	Giordano (1)	
Labbé	All tokens	Instance	Milone (4)	Sereni (2)	Veronesi (1)	
Labbé	All lemmas	Instance	Milone (6)	DeLuca (1)		
NSC	50 MFT, $\Delta = 0.2$	Profile	Balzano (5)	Milone (2)		
NSC	500 MFL, $\Delta = 0.5$	Profile	Veronesi (4)	Milone (1)	Giordano (1)	Naraini (1)
k-NN	50 MFT, Manhattan	Instance	Milone (3)	Murgia (2)	Carofiglio (1)	Balzano (1)
k-NN	1000 MFL, Matusita	Instance	Sereni (5)	Vasta (1)	Raimo (1)	
Bayes	100, Uniform	Profile	Milone (3)	Carofiglio (2)	Balzano (1)	Tamaro (1)
Bayes	500, Prop.	Profile	Milone (4)	Carofiglio (2)	Tamaro (1)	
n-gram	1- and 2-gram, Tani.	Profile	Carofiglio (3)	Mazzucco (3)	Milone (1)	
n-gram	4-gram, Manhattan	Profile	Tamaro (3)	Carofiglio (3)	Milone (1)	

Bayes produces 154 (7 novels × 11 features × 2 uniform prior or proportional). The *k*-NN classifier generates 231 (7 novels × 11 features × 3 distance functions). Finally, the letter *n*-gram model was applied, giving 42 tests (7 novels × 3 distances × 2 uni- and bigrams or 4-grams), resulting in a grand total of 1,519 tests. Some examples are reported in Table 9, where under the label 'First rank' one can find the authors ranked first (in parentheses, we indicate how many times this name appears in first place).

Of the thirty-eight novelists in our corpus, only twenty-two appear, at least once, in first place in the ranked list of authors produced by 1,519 experiments. These names are Balzano, Brizzi, Carofiglio, Covacich, De Luca, De Silva, Giordano, Lagioia, Maraini, Mazzucco, Milone, Murgia, Nesi, Piccolo, Raimo, Rea, Scarpa, Sereni, Tamaro, Valerio, Vasta, and Veronesi. Clearly, instead of having a single alternative author, a larger variability occurs. These experiments also demonstrate that each of the attribution methods focuses on different stylistic aspects and thus proposes different names. This empirically confirms that the underlying methods are independent.

Imposing the additional condition that a name must be the most frequent in the ranked lists generated by an attribution method (with different parameter settings), the list of candidates is limited to six, namely, Balzano (NSC), Carofiglio (*n*-gram), Milone (Labbé, naïve Bayes), Sereni (*k*-NN), Tamaro (*n*-gram), and Veronesi (Delta, NSC). None of them is cited by three (or more) attribution methods. Therefore, the following sixteen names are never strongly associated with Ferrante's style: Affinati, Ammaniti, Bajani, Barricco, Benni, Faletti, Fois, Mazzantini, Montesano, Morazzoni, Nori, Parrella, Pincio, Prisco, Ramondino, and Vinci.

Finally, it is usually useful to inspect some results more carefully. For example, Table 10 depicts the ranked lists established by considering three of Ferrante's novels in the context of the corpus from which Starnone's books have been removed. Compared to Table 3 (same attribution model and feature set), the difference in distance between the first and the second is rather similar to the differences appearing when considering other single rank differences. For example, with the first novel (L'amore molesto), the difference between the first two positions is $0.818-0.812 = 0.006$ (or 0.7%). The gap between the second and the third distance values is $0.843-0.818 = 0.025$ (or 0.3%). As the Delta method must return a ranked list, such a list is generated, but the degree of certainty associated with the first answer is rather low because the latter is too close to the other possible writers. With this parameter setting, and looking at the seven novels by Ferrante, Veronesi appears 3 times in first place, Tamaro twice, and Brizzi and Carofiglio once. All these considerations indicate that the real author is not in the current writer list of thirty-eight candidates.

Table 10 Ranked lists produced by the Delta model after removing Starnone's novels (200 MFT, profile-based approach)

Rank	Distance	Author	Distance	Author	Distance	Author
	L'amore molesto		L'amica geniale		Storia bambina perduta	
1	0.812	Tamaro	0.684	Veronesi	0.706	Veronesi
2	0.818	Brizzi	0.693	Balzano	0.709	Carofiglio
3	0.843	Milone	0.714	Milone	0.756	Tamaro
4	0.861	Lagiola	0.722	Brizzi	0.784	Rea
5	0.880	Balzano	0.734	Nesi	0.789	Balzano

6 Detailed Analysis

When applying the six attribution models, we implicitly admit that the vocabulary choice and the term frequencies can reveal each author's distinctive style. More explicit reasons justifying the strong lexical similarity between Starnone and Ferrante can be found when inspecting the word usage of these two authors, as compared to the others. Focusing on frequent words, one can assume that those words are employed with similar frequencies by all writers. Then, their occurrence frequencies can be compared with the proportion in the novels written by each author. For example, Starnone's books represents 6.4% of the corpus and Ferrante's 6.5%.

Our first example is the word-type padre (father), occurring 9,815 times (100%) in the corpus. Compared to all the other novelists, this word-type is proportionally more frequent in Ferrante's novels (8.5% for 833 occurrences) and in Starnone's writings (11.9% for 1,170 occurrences). A similar distribution can be observed for the word-type madre (mother): its frequency in the corpus is 8,246, with 1,104 in Ferrante's works (13.4%) and 762 in Starnone's (9.2%).

Additional examples can be extracted, and Table 11 reports other word-types such as perciò (therefore) occurring 1,263 times in the entire corpus, with 222 occurrences (17.6%) in Ferrante's novels, and 254 (20.1%) in Starnone's. In the last column of Table 11, the chi-square test has been applied, to verify whether the word-type distribution differs significantly between the authors (all P-values $< 0.1\%$) (Oakes and Farrow, 2007). As a unique case, the word-type persino (even) can also be spelled as perfino. For both Ferrante and Starnone, the preferred spelling is persino (used 266 versus 20 times for perfino in Ferrante's writings, 205 versus 18 times in Starnone's novels). This pattern can also be found in the works of a few other writers, such as Prisco (132 occurrences of persino, 1 of perfino). Some novelists employ only one form (e.g. Baricco with perfino, Tamaro with persino), while others omit both words (e.g. Covacich, Parrella) or use them only rarely (e.g. De Luca or Balzano, with a single occurrence

Table 11 Examples of words occurring more frequently in Ferrante's and Starnone's novels

Word	Corpus	Ferrante (6.5%)	Starnone (6.4%)	Significant?
padre (father)	9,815	833 (8.5%)	1,170 (11.9%)	Yes
madre (mother)	8,246	1,104 (13.4%)	762 (9.2%)	Yes
perciò (therefore)	1,263	222 (17.6%)	254 (20.1%)	Yes
persino (even)	1,351	266 (19.7%)	205 (15.2%)	Yes
temere (fear)	1,345	274 (20.4%)	207 (15.4%)	Yes
tono (tone)	2,135	421 (19.7%)	286 (13.4%)	Yes
gridare (shout)	2,201	399 (18.1%)	303 (13.8%)	Yes
monstrare (to show)	2,271	384 (16.9%)	310 (13.7%)	Yes
content (happy)	1,665	280 (16.8%)	227 (13.6%)	Yes
brutto (ugly)	1,893	327 (17.3%)	243 (12.8%)	Yes
frase (phrase)	2,182	334 (15.3%)	312 (14.3%)	Yes

of perfino). However, this word-type is clearly overused by both Ferrante and Starnone.

On the other hand, some word-types are only employed by these two writers, such as contradditoriamente (fifteen occurrences, contradictory), giravite (thirteen occurrences, screwdriver (more often named cacciavite)), studenti (ten occurrences, students), and soffertamente (eight occurrences, by suffering). An interesting example is the word-type malodore (seventeen occurrences, stink), appearing with this spelling in both Ferrante's and Stanone's novels; however, the same meaning can appear as mal odore or maleodore. These last two spellings appear in other novels but never under Ferrante's or Starnone's authorship.

As a third stratum of word frequency, one can consider word-types with low occurrence frequency in the whole corpus, specifically those occurring more often in works by Starnone and Ferrante than in those by other Italian authors. For example, the term minutamente (minutely) occurs 28 times in Ferrante's novels, 14 times in Starnone's writings, and 3 times in the rest. With tassare (to tax), one can observe something similar: 22 for Ferrante, 10 for Starnone, and 3 times for the others. The word-type reattività (reactivity) occurs 22 times in the whole corpus, with Ferrante employing it 6 times and Starnone 13 times. Our last example relates to dialect usage, with the word strunz (shit). This term does not belong to the classical Italian language (in which it is spelled as stronzo) but corresponds to a Neapolitan dialect form. The distribution of occurrence for this word is as follows: 18 times in Ferrante's novels, 63 times in Starnone's writings, and 4 times for all the others (twice in De Silva's novels, and twice in Raimo's novels).

7 Conclusion

Based on six attribution models and several distinct feature sets, this study confirms the conclusion that Domenico Starnone is the true author of Ferrante's novels (Tuzzi and Cortelazzo, unpublished). Varying the parameter values of the six text categorization methods does not change this conclusion. Considering tokens (Delta, NSC, k-NN, naïve Bayes), lemmas (Delta, Labbé, naïve Bayes), or letter *n*-grams as features, the same result is always achieved. Modifying the feature set size does not change this finding. Applying a classifier based on author profiles (Delta, NSC, letter *n*-grams) or using an instance-based approach (Delta, Labbé, k-NN, naïve Bayes) always produces the same attribution. Thus, considering their lexical proximity, all methods point toward the same name behind Elena Ferrante's novels.

The underlying corpus contains all novelists who have been mentioned as possible hidden hands behind Ferrante. This set contains ten authors originating from the region (Campania) that forms the background for the My Brilliant Friend tetralogy. In addition, when generating this corpus, twelve female writers were selected. Therefore, one can conclude that a real effort has been made to include many authors sharing some important extra-textual relationships with Ferrante (e.g. a woman coming from Naples or its surroundings).

The first part of this study is based on the closed-set assumption. Respecting the underlying hypothesis, we must acknowledge that a collaboration between two (or more) people might exist, for example, to craft some psychological traits of figures appearing in the novels, to give more detail on part of a scene, and to imagine other ways a dialog might take place. However, according to our study, the writing process is the product of a single person. For example, in Table 4 the intertextual distance ranked at #40 is too small to allow for the existence of two writers. The probabilities of assignments derived from both Labbé's method (96%) and the NSC model (99%) are very high, showing strong evidence that only one person undertook the writing. We might also note that Domenico Starnone himself does not corroborate our conclusion (Fontana, 2017), and the mystery about the name Ferrante is not completely erased.

Finally, according to our proposed verification protocol, the Starnone novels were removed from the corpus. Then the same set of methods and feature sets was reapplied. In this open-set situation, each attribution model tends to propose a different name in first place, where the most frequent were

Balzano, Carofiglio, Milone, Sereni, Tamaro, and Veronesi. Such a diverse result indicates that the real author is certainly not present in the list of possible writers. A more detailed analysis performed with the Delta model supports this finding (that the truth lies elsewhere). Finally, the evidence presented in Sections 4, 5, and 6 overwhelmingly points to the same conclusion: Domenico Starnone[2] is the real writer behind the pseudonym Elena Ferrante (even if this conclusion is not 100% certain).

Acknowledgments

The author would like to thank Prof. Arjuna Tuzzi and Prof. Michele Cortelazzo from Padua University for their valuable work in generating the PIC corpus, without which this study could not have been completed. The author also extends his thanks to the anonymous reviewers for their helpful suggestions and remarks.

References

Abbasi, A. and Chen, H. (2008). Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM–Transactions on Information Systems*, **26**(2): Article 7.

Argamon, S. (2008). Interpreting Burrows's Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Burrows, J. F. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. F. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, **22**(1): 27–47.

Chaski, C. (2013). Best practices and admissibility of forensic author identification. *Journal of Law and Policy*, **21**(2): 333–76.

Craig, H. and Kinney, A. F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, **30**(2): 167–82.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, **32**(2): ii4–ii16.

Fontana, E. (2017). Lo scrittore Domenico Starnone: "Io non sono Elena Ferrante". *Il Giornale*, 9 September 2017.

Hoover, D. L. (2004a). Delta prime? *Literary and Linguistic Computing*, **19**(4): 477–95.

Hoover, D. L. (2004b). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Jockers, M. L. (2013). Testing authorship in the personal writings of Joseph Smith using NSC classification. *Digital Scholarship in the Humanities*, **28**(3): 371–81.

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–223.

Jockers, M. L., Witten, D. M., and Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using Delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, **23**(4): 465–491.

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Juola, P. (2016). The Rowling case: a proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, **30**(1): i100–i113.

Juola, P. and Vescovi, D. (2011). Analyzing stylometric approaches to author obfuscation. In Peterson G., Shenoi S. (eds) *Advances in Digital Forensics VII. Digital Forensics 2011. IFIP Advances in Information and Communication Technology*, vol. 361. Berlin: Springer-Verlag, pp. 115–25.

Kocher, M. and Savoy, J. (2017). Distance measures in author profiling. *Information Processing and Management*, **53**(5): 1103–19.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, **14**(1): 33–80.

Labbé, D. (2008). Romain Gary et Emile Ajar. HAL 00279663.

Labbé, C. and Labbé, D. (2013). Duplicate and fake publications in the scientific literature. *Scientometrics*, **94**(1): 379–96.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Mosteller, F. and Wallace, D. L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading: Addison-Wesley.

Oakes, M. P. and Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, **22**(1): 85–99.

Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., and Stein, B. (2017). Overview of PAN'17: author identification, author profiling, and author obfuscation. In Jones G. J. F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappallato, L., and Ferro, N. (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 17)*. Berlin: Springer.

Ratinaud, P. and Marchand, P. (2016). Quelques méthodes pour l'étude des relations entre classification lexicales de corpus hétérogènes: applications aux débats à l'Assemblée nationale et aux sites web de partis politiques. In *Proceedings of JADT*. Jadt.org, pp. 193–202.

Savoy, J. (2012). Authorship attribution based on specific vocabulary. *ACM–Transactions on Information Systems*, **30**(2): 170–99.

Savoy, J. (2013). The federalist papers revisited: a collaborative attribution scheme. In *Proceedings ASIST 2013*, *ASIST*, Montreal, November 2013.

Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, **30**(2): 246–61.

Savoy, J. (2016). Estimating the probability of an authorship attribution. *Journal of the American Society for Information Science and Technology*, **67**(6): 1462–72.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly-discovered Poem? *Biomerika*, **74**(3): 445–55.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(10): 6567–72.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, **18**(1): 104–17.

Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Statistics*, **22**(1): 77–94.

Voorhees, H. and Harman, D. (2005). *The TREC Experiment and Evaluation in Information Retrieval*. Cambridge: The MIT Press.

Witten, I. H., Franck, E., Hall, M., and Pal, C. (2016). *Data Mining. Practical Machine Learning Tools and Techniques*. Amsterdam: Elsevier.

Wolpert, D. H. (2001). The supervised learning no-free-lunch theorems. In *Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications*, Santiago, Chile, September 10–14, pp. 25–42.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Lee, G. G., Yamada, A., Meng, H., and Myaeng, S. H. (eds), *Information Retrieval Technology, Proceedings AIRS 2005*, LNCS, vol. 3689. Berlin: Springer-Verlag, pp. 174–189.

Notes

1 Available at http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

2 The author will give 20 Euros to the first person who provides stronger scientific evidence that the real author behind Ferrante's novels is not Domenico Starnone.