Gender and Genre Linguistic profiling: a case study on female and male journalistic and diary prose

Eleonora Cocciu[•]
Dominique Brunato[◇], **Giulia Venturi**[◇], **Felice Dell'Orletta**[◇]
[•]Università di Pisa
eleonoracocciu.95@gmail.com
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR)
ItaliaNLP Lab - *www.italianlp.it*
{nome.cognome}@ilc.cnr.it

Abstract

English. This paper intends to investigate the linguistic profile of male- and female-authored texts belonging to two very different textual genres: newspaper articles and diary prose. By using a wide set of linguistic features automatically extracted from text and spanning across different levels of linguistic description, from lexicon to syntax, our analysis highlights the peculiarities of the two examined genres and how the genre dimension is influenced by variation depending on author's gender (and vice versa).

Italiano. *Questo lavoro nasce con lo scopo di definire il profilo linguistico di testi scritti da uomini e da donne appartenenti a due generi testuali molto diversi: la prosa giornalistica e le pagine di diario. Attraverso lo studio di una ampia gamma di caratteristiche linguistiche estratte automaticamente dai testi e riguardanti diversi livelli di descrizione linguistica, che vanno dall'analisi lessicale del testo a quella sintattica, questo lavoro mette in luce le peculiarità dei due generi testuali presi in esame e come la dimensione del dominio dei testi venga influenzata dalla dimensione del genere uomo/donna (e viceversa).*

1 Introduction

Authorship profiling is the task of identifying the author of a given text by defining an appropriate characterization of documents that captures the writing style of authors. It is a well-studied area with applications in various fields, such as intelligence and security, forensics, marketing etc. Over the last years, progress in different disciplines such as Artificial Intelligence, Linguistics and Natural Language Processing (NLP) stimulates new research directions in this field leading to the development of 'computational sociolinguistics', a multidisciplinary field whose goal is to study the relationship between language and social groups using computational methods (Nguyen et al., 2016). With this respect, a particular attention has been paid to the influence of gender as a demographic variable on language use. This is a topic that has attracted linguistic research for decades (see e.g. (Lakoff, 1973)) and has received a renewed interest in recent years in the NLP community. The investigation of possible differences between men's and women's linguistic styles has been carried out by using multivariate analyses taking into account gender-preferential stylistic features (Herring and Paolillo, 2006) and machine learning techniques inferring language models that differ at the level of linguistic patterns learned (e.g. based on n-grams of characters, on lexicon, etc.) (Argamon et al., 2003; Sarawgi et al., 2016). These studies have also moved the interest towards the analysis of possible effects driven by textual genres and topics on gender-specific language preferences. With this respect, in the context of the annual PAN evaluation campaign organized since 2013[1], a cross-genre gender identification shared task was newly introduced (Rangel et al., 2016) in 2016, where participants were asked to predict author's gender with respect to a textual typology different from the one used in training. This scenario turned out to be much more challenging for state-of-the art systems, suggesting that females and males can possibly use a different writing style according to genre. While the cross-genre gender prediction task has received attention for many languages, e.g. English, Portuguese, Arabic, the Italian language will be addressed for the first time by the GxG (Gender X-Genre) shared task in the context

[1]https://pan.webis.de/index.html

of the 2018 EVALITA campaign[2].

In line with this interest in the international community, this paper presents a study on gender variation in writing styles with the aim of investigating if there are gender-specific characteristics that are constant across different genres. We define a methodology to carry out an in-depth linguistic analysis to detect differences and similarities in female- and male-authored writings belonging to two different genres. Similarly to the early work by Argamon et al. (2003) for English, our focus is on the linguistic phenomena that contribute to model men's and women's writings in a cross-genre perspective. The main novelty of this work is that we rely on a very wide set of linguistic features automatically extracted from text and capturing lexical, morpho-syntactic and syntactic phenomena. We choose not to focus our analysis on computer-mediated communication texts, which are more typically used in this context, but on two traditional textual genres, i.e. newspaper articles and diary prose.

2 Corpus Collection

The comparative investigation was carried out on two collection of texts, equally divided by gender, and selected to be representative of two different genres: journalistic prose and diary pages.

	Diaries		**Newspapers**	
	Tokens	Document	Tokens	Document
Women	45,155	100	62,469	100
Men	35,493	100	66,860	100
TOTAL	**80,648**	**200**	**129,329**	**200**

Table 1: Corpus internal composition.

For the journalistic genre we collected 200 documents through the advanced search engine available on the website of *La Repubblica*.

For the second textual genre, we collected 200 texts from the website of the Fondazione Archivio Diaristico Nazionale (*National Diaristic Archive Foundation*). In 1984, the Foundation (which is located in Pieve Santo Stefano in the province of Arezzo (Tuscany)) founded a first public archive containing writings of ordinary people, which was changed into the *National Diaristic Archive Foundation* in 1991. Since 2009 the documentary heritage of the archive has been included in the Code of Cultural Heritage of the State.

[2]https://sites.google.com/view/gxg2018

All selected texts were automatically tagged by the part-of-speech tagger described in (Dell'Orletta, 2009) and dependency parsed by the DeSR parser described in (Attardi et al., 2009). Based on the multi–level output of linguistic annotation, we automatically extracted a wide set of more than 170 linguistic features described in the following section.

3 Linguistic Features

Our approach relies on multi-level linguistic features, which were extracted from the corpus morpho-syntactically tagged and dependency-parsed. They range across different levels of linguistic description and they qualify lexical and grammatical characteristics of a text. These features are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability (see e.g. (Biber and Conrad, 2009; Collins-Thompson, 2014; Cimino et al., 2013; Dell'Orletta et al., 2014)).

Raw Text Features: *Token Length* and *Sentence Length* (features 1 and 2 in Table 2): calculated as the average number of characters per tokens and of tokens per sentences.

Number of sentences (feature 3): calculated as the number of sentences of a document.

Lexical Features: *Basic Italian Vocabulary rate features*, all calculated both in terms of lemmata (L) and token (f), referring to *a)* the internal composition of the vocabulary of the text; we took as a reference resource the Basic Italian Vocabulary by De Mauro (2000), including a list of 7000 words highly familiar to native speakers of Italian (feature 4), and *b)* the internal distribution of the occurring basic Italian vocabulary words into the usage classification classes of 'fundamental words', i.e. very frequent words (feature 5), 'high usage words', i.e. frequent words (feature 6) and 'high availability words', i.e. relatively lower frequency words referring to everyday life (feature 7).

Type/Token Ratio: this feature refers to the ratio between the number of lexical types and the number of tokens. Due to its sensitivity to sample size, this feature is computed for text samples of equivalent length, i.e. the first 100 and 200 tokens (feature 8).

Morpho-syntactic Features *Language Model probability of Part-Of-Speech unigrams*: this feature refers to the distribution of unigram

Part-of-Speech (feature 9).
Lexical density: this feature refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text.
Verbal morphology: this feature refers to the distribution of verbs (both main and auxiliary) according to their grammatical person, tense and mood (feature 10).
Syntactic Features *Unconditional probability of dependency types*: this feature refers to the distribution of dependency relations (feature 11).
Subordination features: these features (feature 12) include a) the distribution of subordinate vs main clauses and their average length, b) their relative ordering with respect to the main clause, c) the average depth of 'chains' of embedded subordinate clauses and d) the probability distribution of embedded subordinate clauses 'chains' by depth.
Parse tree depth features: this set of features captures different aspects of the parse tree depth and includes the following measures: a) the depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf (feature 13); b) the average depth of embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers and their distribution of embedded complement 'chains' by depth (feature 14).
Verbal predicates features: this set of features ranges from the number of verbal roots with respect to number of all sentence roots occurring in a text to their arity. The arity of verbal predicates is calculated as the number of instantiated dependency links sharing the same verbal head.
Length of dependency links: the length is measured in terms of the words occurring between the syntactic head and the dependent (feature 15).

4 Data Analysis

For each considered features we calculated the average value and their standard deviation. To investigate which features characterize male vs. female writings, and the possible influence of genre, we assessed the statistical significance of their variation comparing i) male and female writings, independently from the textual genre and ii) diaries and newspaper articles written by women and men. Table 2 reports features that resulted to vary significantly for at least one of the comparisons we considered. In the second and third columns, headed with *Gender*, it is marked the variation with respect to the textual genre, independently from gender's author, the forth and fifth columns, headed with *Genre*, show the statistical significance of variations with respect to gender.

As it can be seen, the number of features that significantly vary is higher in diaries than in newspaper articles (i.e. 23 vs 11); this may suggest that newspapers are characterized by a quite codified writing style with few variations between female and male authors. When we focus on gender, the effect of genre is more prominent for women, as suggested by the greater number of features (i.e. 35) that significantly varies between female diaries and newspaper articles.

Independently from gender, newspapers are characterized by longer words and, among the considered parts-of-speech, by a higher occurrence of prepositions (both simple and articulated), of nouns and proper nouns, as well as by a more extensive use of punctuation. The nominal style characterizing this genre and suggested by the higher proportion of nouns comes out clearly at syntactic level: newspapers articles greatly differ from diary pages since they present a higher percentage of complements modifying a nouns ([11] Compl. and [11] Prep.) also organized in longer embedded chains ([14]), two features which are more common in highly informative texts than in narrative texts like diaries (Biber and Conrad, 2009). According to the literature, these syntactic structures are typically related to sentence complexity as well as deep syntactic trees ([13]) and long clauses ([12] Avg.len.). These phenomena especially distinguish newspaper articles written by men.

As expected, the language of diaries is identified by features typically characterizing narrative texts: the considered collection contains longer sentences, especially male diaries, and a lower percentage of high usage ([6] (f)) and high availability ([7] (f)) lexicon belonging to the *Basic Italian Vocabulary* (BIV). Features capturing the verbal morphology reflect the narrative style used to refer to experiences occurred in the past: the diaries (especially those by male authors) contain a higher usage of imperfect tense and more auxiliary verbs, possibly composing past tenses. In addition, a number of features suggests that the diary

Feature	Gender		Genre		Diaries				Newspaper articles			
	D	J	W	M	Women		Men		Women		Men	
Raw text features												
[1]	-	⋆⋆⋆	⋆	⋆⋆⋆	4.64	(0.31)	4.81	(0.25)	5.07	(0.23)	5.2	(0.22)
[2]	⋆	-	-	⋆	23.95	(20.74)	25.40	(14.53)	25.43	(6.78)	25.49	(6.36)
[3]	-	-	⋆⋆⋆	-	22.16	(14.75)	21.9	(15.61)	26.6	(12.33)	27.8	(11.36)
Lexical features												
[4] (L)	-	-	⋆⋆⋆	-	78.6	(5.44)	72.3	(10.2)	69	(5.47)	68.1	(4.93)
[4] (f)	-	-	⋆⋆⋆	-	88.8	(4.07)	83.9	(6.91)	81.5	(4.00)	80.7	(3.8)
[5] (L)	-	-	⋆⋆⋆	-	83.7	(4.16)	80.2	(4.39)	76.8	(4.14)	76.6	(3.63)
[5] (f)	-	-	⋆⋆⋆	-	81.4	(3.58)	78.9	(3.98)	74.4	(3.93)	74.1	(3.55)
[6] (L)	-	-	⋆⋆⋆	-	11.8	(3.91)	15	(3.84)	17.8	(3.65)	18.3	(3.33)
[6] (f)	⋆⋆⋆	-	-	-	11	(2.52)	12.4	(3.02)	13.9	(2.50)	14.1	(2.36)
[7] (L)	-	-	⋆⋆⋆	-	4.48	(1.85)	4.75	(1.70)	5.42	(1.83)	5.06	(1.68)
[7] (f)	⋆⋆⋆	-	⋆⋆⋆	-	7.55	(2.22)	8.67	(2.53)	11.3	(2.43)	11.8	(2.41)
[8] 100 (f)	-	-	⋆	⋆	0.83	(0.05)	0.83	(0.06)	0.85	(0.05)	0.85	(0.05)
[8] 200 (L)	-	-	⋆	-	0.60	(0.05)	0.61	(0.05)	0.62	(0.04)	0.63	(0.04)
[8] 200 (f)	-	-	⋆⋆⋆	⋆	0.72	(0.05)	0.73	(0.05)	0.75	(0.04)	0.75	(0.04)
Morpho-syntactic features												
[9] Prep.	⋆	⋆⋆⋆	⋆⋆⋆	⋆	11.5	(2.68)	12.6	(2.90)	15.22	(2.12)	16.19	(1.91)
[9] Artic.prep.	⋆	⋆⋆⋆	⋆	⋆⋆⋆	3.27	(1.82)	3.91	(1.53)	5.76	(1.69)	6.50	(1.44)
[9] Pron.	-	-	⋆⋆⋆	⋆	8	(2.79)	7.41	(2.64)	4.37	(1.57)	4.26	(1.21)
[9] Punct.	-	⋆⋆⋆	-	-	13.5	(3.45)	12.6	(3.35)	13.66	(2.42)	12.48	(2.09)
[9] Aux.verb.	⋆⋆⋆	-	-	⋆	2.38	(1.38)	1.80	(1.28)	2.18	(1.52)	2.03	(0.96)
[9] Adj.	-	-	⋆	⋆⋆⋆	4.86	(1.80)	4.89	(1.75)	5.26	(1.58)	5.70	(1.72)
[9] Poss.adj.	⋆	-	-	-	1.46	(0.99)	1.06	(0.86)	0.56	(0.50)	0.60	(0.41)
[9] Neg.adv.	⋆⋆⋆	-	-	⋆⋆⋆	1.68	(1.08)	1.14	(0.65)	0.94	(0.58)	0.85	(0.46)
[9] Subord.conj.	⋆	-	-	-	1.64	(0.92)	1.45	(0.93)	0.95	(0.66)	0.99	(0.54)
[9] Nouns	-	-	⋆⋆⋆	-	19.5	(3.77)	22.8	(4.57)	26.67	(3.36)	26.99	(2.73)
[9] Prop.nouns	⋆	-	⋆⋆⋆	-	2.64	(1.68)	3.70	(3.05)	6.42	(3.11)	6.71	(2.71)
[10] 1p.plur.	⋆	-	-	⋆	4.01	(6.16)	5.35	(8.21)	3.87	(4,74)	2.62	(4.31)
[10] 3p.plur.	-	-	⋆	⋆	14.5	(10.52)	15.5	(12.96)	18.04	(9.17)	18.45	(9.98)
[10] 1p.sing.	⋆	-	⋆	-	20.9	(13.40)	14.5	(12.97)	3.19	(4.41)	2.95	(5.05)
[10] 2p.sing.	-	-	⋆	-	2.80	(5.27)	1.80	(3.45)	0.69	(1.30)	0.45	(1.13)
[10] 3p.sing.	⋆	-	-	⋆	38	(13.28)	45.2	(16.34)	49.64	(13)	50.33	(12.49)
[10] 3p.plur.	-	-	⋆⋆⋆	-	2.31	(3.21)	2.75	(4.50)	6.01	(6.38)	6.34	(5.66)
[10] 1p.sing.	⋆	-	⋆	⋆	7.26	(7.60)	4.32	(6.03)	1.8	(3.91)	0.75	(1.73)
[10] Future	-	-	-	⋆	5.59	(7.40)	2.98	(5.04)	5.94	(8.08)	6.79	(8.95)
[10] Imperfect	⋆	-	⋆⋆⋆	-	21.9	(24.48)	26.2	(24.01)	8.61	(9.10)	9.14	(11.40)
[10] Past	-	-	⋆	-	8.78	(15.17)	9.74	(14.88)	1.51	(4.81)	2.37	(4.70)
Syntactic features												
[11] Compl.	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	-	8.80	(2.15)	9.96	(2.55)	12.10	(1.90)	13	(1.82)
[11] Prep.	⋆⋆⋆	⋆⋆⋆	⋆	⋆	11.5	(2.69)	12.7	(2.88)	15.2	(2.12)	16.2	(1.91)
[11] Punct.	⋆	⋆	⋆	⋆⋆⋆	11.4	(3.05)	10.2	(3)	12.3	(2.22)	11.4	(1.96)
[11] Temp.mod.	⋆	-	⋆⋆⋆	-	0.89	(0.69)	0.61	(0.57)	0.57	(0.43)	0.51	(0.37)
[11] Pred.comp.	⋆	-	-	⋆⋆⋆	2.46	(1.03)	2.03	(1.04)	1.68	(0.70)	1.55	(0.60)
[11] Aux.	⋆	-	-	⋆	2.30	(1.36)	1.72	(1.29)	2.11	(1.56)	1.97	(0.97)
[12] Main	-	-	⋆	⋆⋆⋆	61.1	(14.8)	61.8	(13.7)	67.5	(10.3)	68.1	(10.13)
[12] Sub.	-	-	⋆	⋆⋆⋆	38.9	(14.8)	38.2	(13.7)	32.5	(10.3)	31.9	(10.13)
[12] Avg.len.	⋆⋆⋆	⋆	⋆	-	7.19	(1.17)	7.98	(1.72)	9.20	(1.57)	9.56	(1.46)
[12] (post-verb)	-	⋆	-	-	90.1	(16.9)	87.4	(21.8)	84.2	(13.9)	88.9	(11.06)
[12] (pre-verb)	-	-	⋆⋆⋆	⋆	7.88	(11)	9.56	(15.5)	15.8	(13.9)	11	(11.06)
[13]	⋆	⋆	-	⋆	5.61	(2.84)	6.34	(2.55)	6.21	(1.22)	6.60	(1.18)
[14]	-	⋆	-	-	1.17	(0.12)	1.19	(0.11)	1.29	(0.11)	1.31	(0.08)
[14] (len 3)	-	-	⋆	⋆	1.72	(3.69)	1.68	(2.52)	3.84	(3.14)	3.75	(2.35)
[15]	-	-	⋆⋆⋆	⋆	9.12	(7.47)	9.56	(4.87)	9.84	(2.65)	9.95	(2.66)

Table 2: ⋆⋆⋆ highly statistically significant ($p < 0.001$), ⋆ statistically significant ($p < 0.05$), - any statistically significant features characterizing the two considered textual genres (column *Gender*), i.e. diaries (*D*) vs. newspaper articles (*J*) independently from gender; the two genders (column *Genre*), i.e. women (*W*) vs. men (*M*) independently from textual genre; average feature values and standard deviation in parenthesis for the four different sub-corpora. Features $[1-3]$, [12] Avg.len, [13], [14], [15] are absolute values, the others are percentage distributions.

prose is typically characterized by a more subjective writing style. Namely, the collected diaries present a more extensive use of the first and second singular person verbs, especially those written by women (i.e. 1st person verb: 20.9 women vs 14.5 men), and a higher distribution of possessive adjectives.

If we focus on the gender dimension, our results show that female writings are characterized by features typically found in easier-to-read texts, according to the literature on readability assessment (Collins-Thompson, 2014). This is especially true for the following parameters: they contain shorter words, more fundamental lexicon ([5] (L), (f)), less high usage ([6] (L), (f)) and high availability ([7] (L), (f)) lexicon. At syntactic level, sentences written by women are also characterized by shorter clauses, shorter dependency links and less shallow syntactic trees, as well as by a more canonical use of subordinate clauses in pre-verbal position. On the contrary, men diaries share more features of linguistic complexity: they contain longer sentences, more complex lexicon, a higher percentage of nouns and proper nouns and syntactic features typically occurring in complex structures.

5 Conclusion

We have presented a cross-genre linguistic profiling investigation comparing male and female texts in Italian. We examined a large set of linguistic features, intercepting lexical and syntactic phenomena, which were extracted from two very different textual genres: newspaper articles and diary prose. As expected, the comparative analysis highlighted a number of differences between the two genres, due to the more subjective language characterizing diaries with respect to journalistic prose. Interestingly, we also highlighted that some linguistic features characterize gender dimension and, even more interestingly, we found statistically significant variations also in an objective prose such as newspaper articles.

6 Acknowledgements

The work reported in the paper was partially supported by the 2–year project (2017-2019) UBIMOL, UBIquitous Massive Open Learning, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

S. Argamon, M. Koppel, J. Fine, and A. Shimoni 2003. Gender, Genre, and Writing Style in Formal Written Texts. *Text*, 23(3).

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.

D. Biber and S. Conrad. 2009. Genre, Register, Style. Cambridge: CUP.

A. Cimino, F. Dell'Orletta, G. Venturi, and S. Montemagni. 2013. Linguistic Profiling based on Generalpurpose Features and Native Language Identification. *Proceedings of Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, June 13, pp. 207-215.

K. Collins-Thompson. 2014. Computational Assessment of text readability. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company, 97-135.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.

F. Dell'Orletta, M. Wieling, A. Cimino, G. Venturi, and S. Montemagni. 2014. Assessing the Readability of Sentences: Which Corpora and Features. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, Baltimore, Maryland, USA.

T. De Mauro. 2000. *Grande dizionario italiano dell'uso (GRADIT)*. Torino, UTET.

S. C. Herring and J. C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10/4, pp. 439–459.

R. T. Lakoff. 1973. Language and woman's place. In Language in Society, 2/1, pp. 45–80.

D. Nguyen, A.S. Doruz, C.P. Ros, and F.M.G. de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, Vol. 42, No. 3, Pages 537-593.

F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Pottast, and B. Stein. 2016. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: Balog K., Capellato L., Ferro N., Macdonald C. (Eds.) *CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*. CEUR-WS.org, vol. 1609, pp. 750-784.

R. Sarawgi, K. Gajulapalli, and Y. Choi. 2011. Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, Oregon, USA, June 23-24, 2011, pp. 78–86.