Lost in Text. A Cross-Genre Analysis of Linguistic Phenomena within Text

Chiara Buongiovanni•, Francesco Gracci•, Dominique Brunato◇, Felice Dell'Orletta◇
• University of Pisa
{c.buongiovanni, f.gracci1}@studenti.unipi.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

Moving from the assumption that formal, rather than content features, can be used to detect differences and similarities among textual genres and registers, this paper presents a new approach to the linguistic profiling methodology, which focuses on the internal parts of a text. A case study is presented showing that it is possible to model the degree of variance within texts representative of four traditional genres and two levels of complexity for each.[1]

1 Introduction

The combined use of corpus-based and computational linguistics methods to investigate language variation has become an established line of research. The heart of this research is the so-called 'linguistic profiling', a technique in which a large number of counts of linguistic features automatically extracted from parsed corpora are used as a text profile and can then be compared to average profiles for groups of texts (van Halteren, 2004). Although it has been originally developed for authorship verification and recognition, linguistic profiling has been successfully applied to the study of genre and register variation, following Biber's claim that "linguistic features from all levels function together as underlying dimensions of variation, with each dimension defining a different set of linguistic relations among registers" (Biber, 1993). By modeling the 'form' of a text through large sets of linguistic features extracted from representative corpora, it has been possible not only to enhance automatic classification of genres (Stamatatos et al., 2001), but also to get a better understanding of the impact of features in classifying genres and text varieties (Cimino et al., 2017).

This paper moves in this framework but presents a new approach of linguistic profiling, in which the unit of analysis is not the document as a whole entity, but the internal parts in which it is articulated. In this respect, our perspective is similar to the one proposed by Crossley et al. (2011), who developed a supervised classification method based on linguistically motivated features to discriminate paragraphs with a specific rhetorical purpose within English students' essays. However, differently from that work, we focus on Italian and enlarge the analysis to four traditional textual genres and two levels of language complexity for each. The aim is i) to explore to what extent the internal structure of a text can be modeled via linguistic features automatically extracted from texts and ii) to study whether the variance across different parts of a text changes according to genre and level of complexity within genre.

2 Corpora and approach

Our investigation was carried out on four genres: Journalism, Educational writing, Scientific prose and Narrative. For each genre, we selected the two corpora described in Brunato and Dell'Orletta (2017), which represent a 'complex' and a 'simple' language variety for that genre, where the level of complexity was established according to the expected reader. Specifically, the journalistic genre comprises a corpus of articles published between 2000 and 2005 on the general newspaper *La Repubblica* and a corpus of easy-to-read articles from *Due Parole*, a monthly magazine written in a controlled language for readers with basic literacy skills or mild intellectual disabilities (Piemontese, 1996). The corpus belonging to the Educational genre is articulated into two collections targeting high school (AduEdu) vs. primary school (ChiEdu) students. For the scientific prose,

the 'complex' variety is represented by a corpus of 84 scientific articles on different topics, while the 'simple' one by a corpus of 293 Wikipedia articles, extracted from the Italian Portal 'Ecology and Environment'. For the Narrative genre, we took a dataset specifically developed for research on automatic text simplification. It consists of 56 texts covering short novels for children and pieces of narrative writing for high school L2 students arranged in a parallel fashion, i.e. for each original text a manually simplified version is available. For our study, the original texts and the corresponding simplified versions were chosen as representative of the complex variety and the simple variety, respectively.

All corpora were automatically tagged by the part-of-speech tagger described in Dell'Orletta (2009) and dependency parsed by the DeSR parser (Attardi et al., 2009) to allow the extraction of more than 80 linguistic features, on which we relied to investigate our research questions. These features (detailed in Section 3) capture linguistic phenomena of a different nature, with a focus on morpho–syntactic and syntactic structure, and were selected since they were proven effective for genre classification in previous works, as well as in other scenarios all focused on the analysis of the 'form' of the text rather than its content, such as linguistic complexity, readability assessment (Collins-Thompson, 2014), native language identification (Malmasi et al., 2017).

As a preliminary step for the analyses, all documents were split into a fixed number of sections, where each section is composed by a certain number of paragraphs, roughly corresponding to the three main parts of the rhetorical structure of a text (i.e. introductory, body and concluding paragraphs). According to the literature, for some genres, such as academic writing, the distinction into paragraphs is quite rigid and follows the so-called 'five-paragraphs' format (Crossley et al., 2011) which adheres to the rhetorical goals of the document, i.e. the first and the last paragraph correspond respectively to the introduction and the conclusion, and the three middle ones to the body part. However, based on a preliminary investigation of our corpora we preferred to define a six-section subdivision in order to avoid flattening too much the distinctions across genres. The corpora under analysis indeed are made by documents which are very different in terms of average length: for instance, scientific articles are on average longer than others (184 sentences per document) and this reflects the fact that the body part is more dense and possibly articulated into more middle paragraphs. For each document, the six sections are thus composed by an average number of sentences that depends on the document length, ranging from 2 sentences per section, for the shortest documents, to $\sim$35 for the longest ones. According to this choice, documents shorter than six sentences were discarded, thus we finally relied on a corpus of 1168 documents (see Table 1 for details). As a result of the stage, we represented each section of a document as a vector of features, whose values correspond to the average value that each feature has in all sentences included in the section.

In order to understand whether and to what extent the different parts of a text represent distinctive varieties with a peculiar linguistic structure, we carried out two statistical analyses. First, we assessed whether the difference of the feature values in each section was statistically significant. Specifically, we performed a pairwise comparison between each section and the following one (i.e. 1/2, 2/3, 3/4 etc), as well as between the first and the last section (i.e. 1/6); the latter was deliberately aimed at verifying whether our set of features alone is able to distinguish between the introductory and the closing part of a document, the two more distant sections of a text which are supposed to have a more codified structure. Secondly, we verified whether there is a correlation between the values of features in the two sections under comparison. For both analyses, all data were calculated across and within genre. The cross-genre analysis was focused on genre only, thus considering the two corpora representative of the complex and simple variety as a unique one for each genre. In the second scenario, the two corpora were kept distinct to investigate if there is an effect of genre that is preserved despite language complexity changes.

3 Linguistic features

The set of features extracted from previously identified sections are distinguished into three different categories, according to the level of annotation from which they derive.

Raw Text Features: they include the average word and sentence length (*char_tok* and *n_tokens*

Genre	Corpus	Initial dataset		Analyzed dataset		
		N° Doc	Tokens	N° Doc	Tokens	Avg sentence/section
Journalism	Repubblica (Rep)	318	232.908	304	230.789	5.1
	DueParole (2Par)	321	73.314	303	71.228	2.1
Educational	High-schools educ. materials (AduEdu)	70	48.103	69	47.854	3.9
	Primary schools educ. materials (ChilEdu)	60	23.192	52	22.382	3.5
Scientific Prose	Scientific articles (ScientArt)	84	471.969	84	471.883	35.9
	Wikipedia articles (WikiArt)	293	205.071	249	200.681	4.9
Narrative	Terence&Teacher-original versions (TT orig)	56	27.833	53	25.931	4.2
	Terence&Teacher-simplified versions (TT simp)	56	25.634	54	23.866	4.3

Table 1: Statistics about the corpora used in the study.

in Table 2), calculated as the number of characters per token and of tokens per sentence, respectively.

Morpho-syntactic Features: i.e. distribution of unigrams of part-of-speech distinct into 14 coarse-grained pos tags (cpos_) and the 37 fine-grained tags (pos_) according to the ISST-TANL annotation.

Syntactic Features: these features model grammatical phenomena of different types, i.e:
- the *probability of syntactic dependency types* e.g. subject (*dep_subj*), direct object (*dep_dobj*), modifiers, calculated as the distribution of each type out of the total dependency types according to the ISST-TANL dependency tagset;
- the *length of dependency links*, i.e. the average length of all dependency links (each one calculated as the number of words occurring between the syntactic head and the dependent) (*avg_links_l*) and of the maximum dependency link (*max_links_l*);
- the *order of constituents* with respect to the syntactic head: as a proxy of canonicity effects, it is calculated the relative position of the subject, object and adverb with respect to the verbal head and the position of the adjective with respect to the nominal head;
- the *parse tree structure*, in terms of features calculating: the depth of the whole parse tree (*sent_depth*) (in terms of the longest path from the root of the dependency tree to some leaf); the width of the parse tree (*sent_width*), measured as the highest number of nodes placed on the same level; the average number of dependents for all verbal and nominal heads (*avg_dependent*);
- *subordination features*: within the group of syntactic features, a in–depth analysis was devoted to model subordination phenomena by measuring: the average distribution of subordinate clauses for sentence (*avg_sub_clause*), the percentage of subordinate clauses with respect to the main clause (*% sub_main*) and the percentage of embedded subordinate clauses, i.e. subordinate clauses dependent on other embedded subordinate clauses (*% sub_minor*); for each type, it is also calculated the average depth (*subord_depth*) and weight (*subord_width*) of the parse tree generated by the subordinate clauses and their relative order with respect to the clause on which they depend.

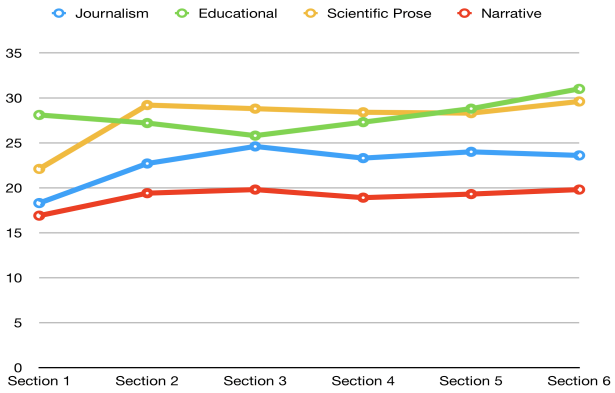

Figure 1: Average sentence length in the 6 sections across genres.

4 Data Analysis

Table 2 illustrates the main findings we obtained. Specifically, it shows all features which turned out to have a statistically significant variation in at least one of the six pairwise comparisons, or a correlation score > 0.3 according to the Spearman's correlation coefficient. A first clear result is that the higher number of features varying in a statistically significant way occurs in the journalistic and scientific genre, both considered as whole (i.e. row *g* for each feature) and with respect to the language complexity variety (rows *s* and *c*). The opposite trend is reported for educational texts, which is probably due to the heterogeneous nature of this

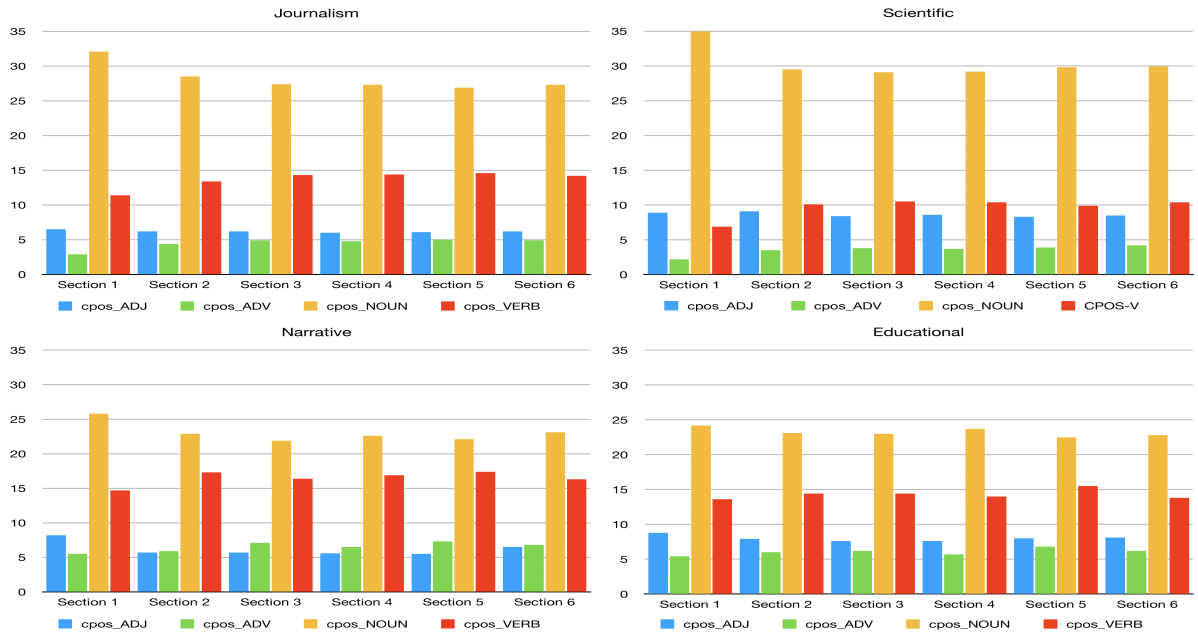

Figure 2: Distribution of lexical parts-of-speech in the four genres.

genre that includes documents of different textual typologies (course books, pieces of literature etc.).

If journalism and scientific prose are the two genres with the highest internal variance, the comparison between sections allows us to get a better understanding of this data. Specifically, for both genres, the majority of significant variations are observed between the first and the second section and between the first and the last one. This suggests that the introduction is a stylistic unit with a peculiar linguistic structure with respect to the body and the conclusion. It is characterized e.g. by shorter sentences (Figure 1), likely due to the presence of the title in both newspaper and scientific articles, and by a distinctive distribution of Parts–of–speech (Figure 2). With this respect, this data are consistent with other studies in the literature, e.g. (Voghera, 2005), and also with previous findings we obtained on the same corpora (Brunato et al., 2016), showing that scientific prose and newswire texts rely more on the nominal style. However, with the proposed approach, we were able to go further in this analysis, highlighting that noun/verb ratio is always higher in the first section than all other ones. Besides, at least for newspaper articles, this feature appears as a genre marker which is not affected by language complexity, since the same tendency is observed when the 'simple' and the 'complex' corpus are analyzed independently. The same does not hold for other features related to syntax and, in particular, to the use of subordination. In this case, the 'shift' between the introduction and the subsequent part of texts yields significant variations only for articles of *Repubblica*. Specifically, the first section contains less embedded sentences (*sent_depth*: 1st sect: 5.55; 2nd sect: 7.76), and a lower presence of subordinate clauses, which appear as structurally simpler e.g. in terms of depth (*subord_depth*: 1st sect: 1.67; 2nd sect: 3.5) and width (*subord_width*: 1st sect: 0.94; 2nd sect: 1.97). Conversely, for the simple variant of this genre (i.e. the articles of the easy-to-read newspaper *2Parole*), we do not observe significant changes affecting these features; this is not particularly surprising since subordination is always less represented in this corpus with respect to all the other ones.

Leaving aside the similar tendencies characterizing the introduction, Journalistic and Scientific prose show a different behavior when we focus on the internal structure of text. While in this case much fewer features vary in a significant way, the majority occurs in the journalistic genre only, especially between the second and the third section. Again, they concern a different distribution of morpho-syntactic categories but also some syntactic features related to subordination. According to these data, we can conclude that the journalistic genre has a more rigorous structure and that it is possible to capture the boundaries between different parts by using linguistic features that are not related to the content of the article.

features		Journalism						Scientific Prose						Narrative						Educational					
		1/2	2/3	3/4	4/5	5/6	1/6	1/2	2/3	3/4	4/5	5/6	1/6	1/2	2/3	3/4	4/5	5/6	1/6	1/2	2/3	3/4	4/5	5/6	1/6
Raw text features																									
n_tokens	g	✓✓	✓*	-*	-*	-*	✓✓	✓✓	-	-	-*	-*	✓✓	-	-	-	-	-	✓*	-*	-*	-*	-*	-*	-*
	s	✓✓	-	-	-	-	-	✓✓	-	-	-	-	✓✓	-*	-	-	-	-	✓	-*	-*	-*	-*	✓*	-*
	c	✓✓	-	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	✓	-*	-	-*	-	-	-
char_tok	g	-*	-*	-*	-*	-*	✓	✓	-	-*	-*	-	✓	-	-*	-*	-	-	-	-*	-*	-*	-*	-*	-*
	s	-*	-*	-*	-*	-	✓✓	✓	-	-*	-	-	✓✓	-	-*	-*	-*	-	-	✓*	-*	-*	-*	-*	-*
	c	-	-*	-*	-*	-*	-	-*	✓*	-*	-*	-*	-*	-	-*	-*	-	-	-	-*	-*	-*	-*	-*	-*
Morpho-syntactic features																									
cpos_ADJ	g	-	-	-	-	-	-	✓	✓	-*	-	-	*	-	-	-	-	-	-	-	-*	-*	-	-*	-*
	s	✓	-	-	-	-	-	✓✓	-	-*	-	-	✓	-	-	-	-	-	-	✓✓	-	-	-	-	-*
	c	✓	-	-	-	-	✓	-	-*	-*	-*	-*	-*	-	-	-	-*	-*	-*	-	-*	-*	-	-*	-
cpos_ADV	g	✓✓*	✓*	-	-	-	✓✓	✓✓*	-	-	-*	-	✓✓	-	-	-	-	-	✓	-	-	-*	✓	-	-
	s	✓✓	-	-	-	-	✓✓	✓✓*	-	-	-	-	✓✓	-	-	-	-	-	✓	✓	-	-*	✓	-	-
	c	✓✓*	-*	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-	-	-	-*	-	-	-
cpos_CONJ	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	✓	-	-	-	✓	✓	-	-	-
	s	✓✓	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	✓	-	-	-	-
	c	✓✓	-	-	-	-	✓✓	✓*	-*	-*	-*	-*	✓*	-	-*	-	-	-*	-	-	-	-	-	-	-
cpos_NOUN	g	✓✓*	✓*	-*	-*	-*	✓✓	✓✓	-	-	-*	-	✓✓	✓✓	-	-	-*	-	✓✓*	-	-*	-*	-*	-*	-*
	s	✓✓*	-*	-	-	-*	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-*	-	✓✓	✓	-*	-*	-	-	-*
	c	✓✓*	✓✓*	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	✓*	-	-	-*	-	✓*	-	-	-*	-*	-*	-
pos_PROP_N	g	✓✓*	-*	-*	-*	-*	✓✓*	✓✓	-*	-*	✓*	-*	✓✓	✓	-*	-*	-*	-*	-*	-*	-*	-*	-*	-*	-*
	s	✓✓*	-*	-*	-*	-*	✓✓*	✓✓	-*	-*	-*	-*	✓✓	-	-*	-*	-*	-*	-	✓*	-*	-*	-*	-*	-*
	c	✓✓	✓✓	-	-*	-	✓✓	-*	-*	-*	-*	-*	-*	-	-*	-*	-*	-*	-	-*	-*	-*	-*	-	-*
cpos_VERB	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-	-	-	-	-	-	✓	✓	-
	s	✓✓	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-	-	-	-	-	-	✓	✓✓	-
	c	✓✓	✓✓	-	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	✓	-	-	-	-	-	-	-*	-*	-	-	-
pos_AUX	g	✓*	✓*	-*	-*	-*	✓	✓✓*	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-	-	✓	-
	s	-*	-*	-*	-*	-*	-	✓✓	-	-	-	-	✓✓	-	-	-*	-	-	-	-	-	-	-	-	-*
	c	✓✓*	✓✓*	-*	-*	-*	✓✓	-*	-	-	-*	-	-*	-	-	-	-*	-	-	-	-	-	-	-	-
Syntactic features																									
dep_dobj	g	✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	✓	✓	-	-	✓	-	✓	-	-	-	✓
	s	-	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	✓✓	-*	-	-	✓	-	✓✓	-	-*	-	✓
	c	✓✓	-	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-	-	-	-	-	-	-
dep_subj	g	-	-	-	-	-	-	✓✓	-	-	-	-	✓✓	✓	-	-	-	-	-	-	-*	-	-	-	-
	s	-	-	-	-	-	-	✓✓	-	-	-	-	✓✓	-	-*	-	-	-	-	-*	-*	-	-*	-	-
	c	✓	-	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	✓	-	-	-	-	-	-	-	-	-	-	-
max_links_l	g	✓✓	-	✓*	-*	-	✓✓	✓✓	-	-	-*	-	✓✓	-	-	-	-	-	-*	-	-*	-*	-*	-*	-*
	s	✓✓	-	-	-	-	✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-*	-	-*	-*
	c	✓✓	-	✓	-*	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-*	-	-*	✓	-	-	-
avg_links_l	g	✓✓	-	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-	-	-*	-
	s	✓	-	-	-	-	-	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-*	-	-*	-
	c	✓✓	-	✓	-	-	✓✓	-*	-*	-*	-*	-*	-	-	-	-*	-	-	-*	-	-*	-	-	-	-
sent_depth	g	✓✓	-*	-*	-*	-*	✓✓	✓✓	-	-	-	✓*	✓✓	-*	-	-*	-*	-*	-	-*	-*	-*	-*	-*	-*
	s	-	-	-	✓	-	✓	✓✓	-	-	✓	-*	✓✓	-*	-	-	-*	-*	-	-	-*	-*	-*	-*	✓*
	c	✓✓	-	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-	-*	-*	-	-	-*	-*	-	-*	-*	-*
sent_width	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-*	-*	-	-	-*	-
	s	✓✓	-	-	-	-	-	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-*	-*	-	-	-*	-*
	c	✓✓	-	✓	-*	-	✓✓	-*	-*	-*	-*	-*	-*	-	-*	-	-	-	-*	-	-*	✓	-*	-	-
avg_dependent	g	✓✓	✓	-*	-*	-*	✓✓	✓✓	-*	-*	-*	✓*	✓✓	✓	-*	-*	-	-*	✓	-*	-*	-*	-*	-*	-*
	s	✓	-	-	-	-*	-	✓✓	-	-	-	-	✓✓	-	-	-*	-	-*	✓	-	-*	-*	-*	-*	-*
	c	✓✓	✓	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-*	-*	-	-*	✓	-*	-*	✓	-*	-*	-*
Subordination features																									
avg_sub_clause	g	✓✓	✓*	-*	-	-*	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-*	-*	✓✓	-	-	-	-	-*	-*
	s	-	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-*	-*	✓	-	-	-	-	-*	-*
	c	✓✓	-	-	-*	-	✓✓	-*	-*	-	-*	-*	-*	✓✓	✓	-	-	-	-	-*	-	-	-	-	-
subord_depth	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	✓	✓✓	-*	-	-*	-*	-*	-	-	-	-*	-*	-	-
	s	-	-	-	-	-	✓	✓✓	-	-	-	-	✓✓	-	-	-	-*	-*	✓✓	-	-	-*	-*	-*	✓
	c	✓✓	-	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-*	-	-	-	✓	-	-	-	-
subord_width	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	✓	✓✓	-	-	-	-	-	-	-	-	-*	-	-	-
	s	-	-	-	-	-	✓	✓✓	-	-	-	-	✓✓	-	✓	-	-*	-*	✓	-	✓	-*	-*	-*	-
	c	✓✓	-	-	-	-	✓✓	-*	-*	-	-*	-*	-*	-	-	-	-*	-	-	-	-	-	-	-*	-
% sub_main	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-*	✓✓	-	-	-	-	-	-
	s	-	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-*	✓✓	-	✓*	✓✓*	-*	✓*	-
	c	✓✓	-	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-	-	✓	-	-	-	-
% sub_minor	g	✓✓	-*	-	-*	-	✓✓	-	-	-	-	-	✓	-	-	-	-	✓	-	-	-*	✓	-	-	✓
	s	-	-	-	-	-	✓	-	-	-	-	-	✓	-	-	-	-	-*	-*	-	-	✓	-	-*	-*
	c	✓✓	-	-	-	-	✓✓	-*	-*	-	-*	-	-	✓	-	-	-*	-	-	-	-	-	-*	-	✓

Table 2: A set of linguistic features resulting as significant in at least one pairwise comparison. ✓✓ means highly statistically significant ($p < 0.001$), ✓ statistically significant ($p < 0.05$), - no significance; * correlation related to the Spearman's rank correlation coefficient (rho > 0,3), g=global corpus, s=simple variety of the corpus, c=complex variety of the corpus.

5 Conclusion

In this paper we have presented a novel approach to the study of language variation, which relies on the prerequisites of the linguistic profiling methodology but with the specific purpose of modeling the stylistic form of the different parts within a text. A cross-genre investigation on four traditional genres in Italian, and two levels of complexity for each, showed that morpho-syntactic and syntactic features are differently distributed across subsections of texts belonging to a specific genre and language variety. This approach has important implications for research on genre variation since it suggests that the characterization of texts and texts varieties should benefit by inspecting corpora from this fine-grained perspective. A better understanding of linguistic phenomena characterizing the introductory, middle and conclusive parts of a text is also highly relevant not only to enhance automatic genre classification but also for other natural language processing applications devoted to modeling style: e.g. in education, as a component of intelligent tutoring systems able to provide detailed feedback to students in writing courses or for the automatic generation of texts with the stylistic properties of a specific genre and level of complexity.

Acknowledgments

This work was partially supported by the 2-year project ADA, Automatic Data and documents Analysis to enhance human-based processes, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Joseph 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.

Douglas Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2), 219–242.

Dominique Brunato and Felice Dell'Orletta. 2017. On the order of words in Italian: a study on genre vs complexity. *International Conference on Dependency Linguistics (Depling 2017)*, 18-20 September 2017, Pisa, Italy.

Dominique Brunato, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. 2016. Monitoraggio linguistico di Scritture Brevi: aspetti metodologici e primi risultati. *A. Manco e A. Mancini (eds.), Scritture Brevi: segni, testi e contesti. Dalle iscrizioni antiche ai tweet, Collana di studi Quaderni di AION-Linguistica, Università di Studi di Napoli "L'Orientale"*, Napoli, 149–176.

Andrea Cimino, Martijn Wieling, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi. 2017. Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax. *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, 11-13 December, 2017, Rome.

Kevyn 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.

Crossley, S.A., Dempsey, K., McNamara, D.S. 2011. Classifying paragraph types using linguistic features: Is paragraph positioning important? *Journal of Writing Research*.

Felice 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.

S. Malmasi, E. Keelan, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*.

Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata.* Napoli, Tecnodid.

Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis. 2001. Automatic text categorization in terms of genre and author. Computational Linguistics, (26) 471–495.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In Proceedings of the Association for Computational Linguistics (ACL04), 200207.

Miriam Voghera. 2005. La misura delle categorie sintattiche. In Chiari Isabella / De Mauro Tullio (eds.) Parole e numeri. Analisi quantitative dei fatti di lingua, Aracne, Roma, 125–138.