ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 6, Number 1
june 2020

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn PDF 9791280136404

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_6_1

IJCoL

Volume 6, Number 1
june 2020

CONTENTS

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

Chiara Buongiovanni*
Università di Pisa

Francesco Gracci**
Università di Pisa

Dominique Brunato†
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab

Felice Dell'Orletta‡
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab

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 linguistic profiling – a well-established methodological framework to study language variation – which is applied to detect significant variations within the internal structure of a text. We test this approach on the Italian language using a wide spectrum of linguistic features automatically extracted from parsed corpora representative of four main genres and two levels of complexity for each, and we show that it is possible to model the degree of stylistic variance within texts according to genre and language complexity.

1. Introduction

The combination of corpus-driven and Natural Language Processing (NLP)-based approaches to study language variation and use from a stylistic and sociolinguistic perspective has become an established line of research. The heart of this research is the so-called 'linguistic profiling', a framework of analysis in which a large number of counts of linguistic features extracted from linguistically annotated corpora are used as a text profile and can then be compared to average profiles of texts (or groups of texts) to identify those that are similar, at least similar in terms of the profiled features (van Halteren 2004; Montemagni 2013). Although it has been originally developed for authorship verification and recognition, i.e. to select one author from a set of known authors or to confirm or deny authorship by a single known author, this methodology has been successfully applied to research on genre and register variation, as well as in a variety of scenarios focused on modeling the 'form' rather than the content of texts: from the identification of developmental patterns in typical (Lu 2009; Lubetich and Sagae 2014) and atypical language acquisition (Prud'hommeaux et al. 2011; Rouhizadeh,

* Dipartimento di Filologia Letteratura Linguistica Piazza Torricelli, 2, 56126, Pisa.
E-mail: c.buongiovanni@studenti.unipi.it
** Dipartimento di Filologia Letteratura Linguistica Piazza Torricelli, 2, 56126 Pisa.
E-mail: f.gracci1@studenti.unipi.it
† ItaliaNLP Lab (www.italianlp.it), ILC–CNR - Via Moruzzi, 1 - 56124, Pisa, Italy.
E-mail: dominique.brunato@ilc.cnr.it
‡ ItaliaNLP Lab (www.italianlp.it), ILC–CNR - Via Moruzzi, 1 - 56124, Pisa, Italy.
E-mail: felice.dellorletta@ilc.cnr.it

Sproat, and van Santen 2015) and in school learners' writing (Barbagli et al. 2015), to the detection of linguistic markers of adult cognitive impairments (Roark, Mitchell, and Hollingshead 2007); besides, from a computational sociolinguistics perspective, for studying variations related to the social dimension of language (Nguyen et al. 2016) or for modeling stylometric characteristics of authors or author groups (Daelemans 2013).

The assumptions of linguistic profiling lay in the framework of the Multi-Dimensional Analysis (MDA) pioneered by Douglas Biber, a text-linguistic approach to characterize language use across social and communicative contexts through the quantitative and functional analysis of co-occurrence patterns of linguistic features and underlying dimensions of language use (Biber, Conrad, and Reppen 1998). The important hallmark of MDA is 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). This set of features should capture properties of text that are 'signatures' of its style rather than the topic it deals with, since they are more appropriate to investigate differences and similarities between texts from a functional perspective. Since its foundation, MDA has also aimed to be "computer-based in that it depends on automated analyses of linguistic features in texts. This characteristic enables distributional analysis of many linguistic features across many texts and text varieties". However, as Argamon observes in its recent survey, traditional research in computational register analysis has exploited a limited number of stylistic features, such as the relative frequencies of function words, taken as indicative of different grammatical choices rather than topic ones, or of character n-grams assumed to model linguistic variation in lexical, grammatical, and orthographic preferences (Argamon 2019). This is easily explainable as these features are simpler to extract from text than more sophisticated syntactic and discourse features, yet effective enough to capture dimension of variations. In fact, if function words require language-specific lists of a few hundred words (such as pronouns, prepositions, auxiliary and modal verbs, conjunctions, and determiners), character n-grams do not need any form of language-specific preprocessing, not even tokenization, which instead might be necessary for word n-grams.

More recently, the development of robust and fairly accurate NLP pipelines together with the increased computational power to process large volumes of data has allowed to automatize the process of feature extraction from large-scale corpora, enhancing the potential contribution of linguistic profiling for studying language variation. By modeling the 'form' of a text through large sets of features spanning across distinct levels of language description, it has been possible not only to improve automatic classification of genres (Stamatatos, Fakotakis, and Kokkinakis 2001), but also to get a better understanding of the impact of those features in classifying genres and text varieties (Cimino et al. 2017; Finn and Kushmerick 2006). This paper adopts this framework but, quite differently from much previous research, it presents a new application 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. A close perspective has been pursued by (Crossley, Dempsey, and McNamara 2011) but with a different aim, i.e. the automatic discrimination of paragraphs with a specific rhetorical purpose in English students' essays by using a feature engineering approach. In our contribution, we focus on the Italian language and we broaden the scope of the analysis to four traditional textual genres and two levels of linguistic complexity for each. Briefly, this paper intends to answer the following research questions:

i) to what extent is NLP–based linguistic profiling a viable approach to characterize the internal structure of a text?

ii) does the variance across different parts of a text change according to genre and level of complexity?

In what follows we first present the corpora on which the study was carried out and then move to the presentation of the approach, together with the description of the linguistic features used for analysis. In Section 5, we discuss the main findings we obtained and outline the primary conclusions.

2. Corpora

Our investigation was carried out on four traditional genres: Journalism, Educational material, 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 relied on a dataset specifically developed for research on automatic text simplification. It consists of 56 short novels for children and pieces of narrative writing for L2 high school students arranged in a parallel fashion, i.e. for each original text a manually simplified version is available. For our investigation, the original texts and the corresponding simplified versions were chosen as representative of the complex variety and the simple variety of Narrative, respectively.

All these corpora contain documents which are very different in terms of length: for instance, scientific articles are on average longer than others (215 sentences per document) and this reflects the fact that the body part is more dense and possibly articulated into more middle paragraphs. On the contrary, the easy–to–read newspaper articles (i.e. *Due Parole*) are made of $\sim$12 sentences. Thus, for each document, we also defined an internal subdivision into six parts intended to enable the linguistic profiling investigation on the internal structure of text. The rationale of the splitting approach is outlined in the following section. As a result, we ended up with six sections per document, where each section is composed by a number of sentences that depends on the average document length, ranging from two sentences per section for the shortest documents, to $\sim$35 for the longest ones. Moreover, according to this approach, documents shorter than six sentences were discarded. As a result of the whole process, the final corpus is made of 1,168 documents. Table 1 reports general statistics about the final dataset in terms of: number of documents in each corpus, corpus size in number of tokens, average number of sentences per document and average number of sentences per section in each corpus.

3. Methodology

As a first step we focus on inspecting differences and similarities across all corpora considering the document as the unit of analysis. This is a more traditional perspective

Table 1
Corpus statistics ("c" and "s" stand for 'complex' and 'simple' variety for each genre.)

Genre		Corpus	N. doc	Tokens	Sent/doc	Sent/sect
Educational	c	High-school texts (*AduEdu*)	69	47.854	23.4	3.9
	s	Primary school texts (*ChilEdu*)	52	22.382	21.0	3.5
Journalism	c	la Repubblica articles (*Rep*)	304	230.789	30.5	5.1
	s	Due Parole articles (*2Par*)	303	71.228	12.7	2.1
Narrative	c	Terence&Teacher original (*TT orig*)	53	25.931	25.3	4.2
	s	Terence&Teacher simplified (*TT simp*)	54	23.866	25.5	4.3
Scientific prose	c	Scientific articles (*ScientArt*)	84	471.883	215.6	35.9
	s	Wikipedia articles (*WikiArt*)	249	200.681	29.1	4.9

to address linguistic profiling and it is useful to provide quantitative and quality data on the average variation across the examined genres and internal level of complexity.

To carry out this analysis all corpora were firstly automatically tagged by the part-of-speech tagger described in (Dell'Orletta 2009) and dependency parsed by the DeSR parser (Attardi et al. 2009). DeSR, trained on the ISST–TANL treebank (Montemagni et al. 2003) consisting of articles from newspapers and periodicals, achieves a performance of 83.38% and 87.71% in terms of Labeled (LAS) and Unlabeled Attachment Score (UAS) respectively, when tested on texts of the same type. However, it is widely acknowledged that even state–of–the art parsers have a drop of accuracy when tested against corpora differing from the typology of texts on which they were trained (Gildea 2001). Therefore, we can assume that the performance of DeSR would be possibly lower when parsing texts of a different textual genre, such as narrative or scientific writing. Despite this fact, we also expect that the distributions of parsing errors will be almost similar, at least when analysing texts of the same domain and language variety, thus allowing us to carry out an internal comparison with respect to examined linguistic parameters. Besides, the effect of genre variation on the performance of general–purpose parsers is likely to be less strong since all genres here considered contain *standard* texts, i.e. texts linguistically similar to the ones used in training.

Based on the output of the different levels of linguistic annotation, we automatically extract from text a wide set of linguistic features, thus creating a feature-vector representation of each document, where each dimension of the vector corresponds to the average value of a given linguistic feature in the document. The set of features used in our study, and the motivation underlying their choice, are described in the following section.

For what concerns the second analysis, which is the most innovative perspective of this study, we looked at the internal structure of documents and we investigate how the same linguistic features vary across different sections of text. To allow this investigation we firstly defined the 'new' unit of analysis as follows. 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 previous work, for some genres such as academic writing, the distinction into paragraphs is quite rigid and follows the so-called 'five-paragraphs' format (Crossley, Dempsey, and McNamara 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 showing that the average document length is highly variable, we preferred to define a six-section subdivision in order to avoid flattening too much the distinctions across genres.

Finally, we performed the feature extraction process, this time representing each section of a document as a vector of features, whose values correspond to the average value that each linguistic feature has in all sentences included in the section.

To understand whether and to what extent the different sections of a text represent distinctive varieties with a peculiar linguistic structure, we carried out two statistical analyses. The first one aimed to assess the significance of variation between the same features extracted from different sections. 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 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 most distant sections of a text which are supposed to have a more codified structure. Secondly, we evaluated 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 collapsing the internal distinction in terms of complexity and considering the two corpora as a unique one for each genre. In the within-genre condition, the two corpora were kept distinct thus allowing us to observe whether there is an effect of genre that is preserved despite changes in linguistic complexity.

4. Linguistic Features

The set of features used for our analysis models a variety of phenomena related to the sentence structure, with a particular focus on morpho–syntactic and syntactic properties. These features were selected since they proved to be effective predictors of systematic variations in automatic genre classification (Cimino et al. 2017), as well as in other tasks in which the 'form' of the text is more relevant than the content, such as the prediction of perceived sentence complexity by humans (Brunato et al. 2018), the assessment of text readability (Collins-Thompson 2014) or the identification of the native language of speakers from their productions in a second language (Malmasi et al. 2017).

The features can be 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 (*word_length* and *sent_length* in Tables 2 and Tables 3), 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 distinguished into 14 coarse-grained pos tags (*cpos_*) and the 37 fine-grained tags (*pos_*) according to the ISST-TANL tagset.

Syntactic Features: features modeling syntactic phenomena of different types, i.e.:
- the *distribution of syntactic dependency types* (*dep_*), e.g. subject, direct object, modifiers, calculated as the distribution of each typed out of the total dependency relations according to the ISST-TANL dependency tagset;
- the *length of dependency links*, i.e. the average length of all dependency links (*avg_links_l*) and of the longest link (*max_links_l*). For each link, the distance is calculated as the number of words occurring between the syntactic head and the dependent;

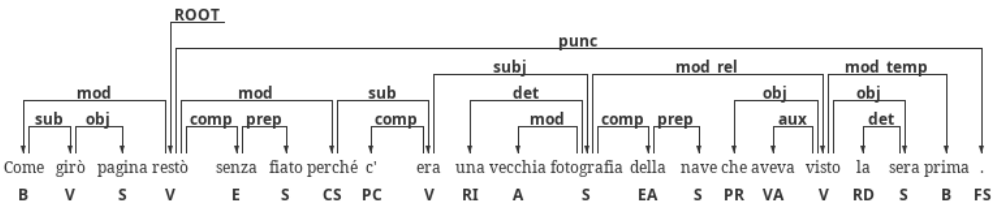

Figure 1
Graphical representation of the annotated sentence in (1).

- the *order of constituents* with respect to the relative lexical head: this feature works as a proxy of canonicity effects and it is calculated for the main elements of the clause in terms of: the distribution of pre-verbal and post-verbal subjects (*pre_subj*, *post_subj*), of pre-verbal and post-verbal object (*pre_obj*, *post_obj*), of pre-nominal and post-nominal adjectives (*pre_adj*, *post_adj*), and of pre-verbal and post-verbal adverbs (*pre_adv*, *post_adv*);
- the *parse tree structure*, in terms of features corresponding to: the average depth of the whole parse tree (*parse_depth*) (i.e. the longest path from the root of the dependency tree to a leaf); the average width of the parse tree (*parse_width*), where the width is measured as the average number of nodes placed on the same level; the average number of dependents for all heads in the sentence (*avg_dep_all*), for the verbal heads and for the nominal heads (*avg_dep_verb*, *avg_dep_noun*);
- *subordination features*: a thorough analysis was devoted to investigate the use of subordination by computing: the average distribution of subordinate clauses with respect to the main clause (*sub_main*) and of embedded subordinate clauses (i.e. subordinate clauses dependent on other subordinate clauses) out of all subordinate clauses (*sub_minor*). In addition, both for the 'superordinate' subordinate clause and the embedded ones, it is calculated the relative order with respect to the clause on which they depend (*pre_sub, post_sub*), as well as the average depth (*subord_depth*) and the average width (*subord_width*) of the parse tree generated by the subordinate clause.

To exemplify some of the above-described features, we refer to the following sentence extracted from the Narrative corpus, whose graphical visualization is reported in Figure 1.

(1) Come girò pagina restò senza fiato perché c'era una vecchia fotografia della nave che aveva visto la sera prima.

The sentence has a length of 21 tokens (punctuation included) and the average word length is 4.62 characters. For what concerns the POS distribution, it presents e.g. an equal distribution of verbs and nouns, each one corresponding to the 23.81% of the whole POS. With respect to syntax–related features, the average length of dependency links is 1.84 and the maximum link (excluded punctuation) is 5–token long (corresponding to the *mod_rel* dependency going from the nominal head 'fotografia' to the embedded verb of the relative clause 'visto'). The maximum parse tree depth is 6, corresponding to the longer path, in terms of number of intervening nodes, from the root of the sentence ('restò') to the more distant leaf, the definite article 'la'.

5. Data Analysis

5.1 Differences and similarities across genres and complexity

Table 2 reports the average value and standard deviation of all examined linguistic features in the four corpora considered as a whole, as well as considering the 'simple' and 'complex' sub–corpora as distinct varieties. Along with the mean and standard deviation value, we also compute the coefficient of variation, which corresponds to the ratio between the standard deviation and the mean value of the feature. This measure allows us to evaluate the dispersion of values around the average in a standardized way, that is comparing the stability of features pertaining to data measured on different scales. The assumption in exploiting this metric is that the more stable a feature in a given corpus, the more meaningful it is for characterizing it. Of course, we expect that features with the lowest coefficient of variation identify general tendencies of a language, in this case of Italian. However, the effect of genre and language variety can be inferred by the different positions that these 'most stable' features have in the ranking of each corpus. With this respect, in Table 2 we also indicate the first ten features ranked according to the coefficient of variation for each corpus.

A first comparison across corpora based on the average feature value shows some expected tendencies, such as the highest sentence and word length characterizing scientific prose, also in the simple variety. Both these features are considered as raw indicators of sentence complexity and they are used by traditional indexes to evaluate the readability of a text (Collins-Thompson 2014). Focusing on the relative frequencies of core parts of speech, which is traditionally used as a marker for text genres (Biber, Conrad, and Reppen 1998), we can also note that scientific and journalistic texts use more nouns and less verbs in comparison to other genres, a marker of the nominal style featuring these texts. The opposite holds true for the narrative prose, which indeed has the lowest noun/verb ratio. Note that with the only exception of sentence length, all these parameters occur among the ten most stable ones in almost all genres and varieties, with the average word length always in the top two positions. The distribution of nouns is much prone to variation, especially in Narrative.

If we consider features modeling the syntactic profile of text, we can observe that Scientific articles again display values typically reported in readability studies to describe difficult-to-read texts, such as deeper parse tree. If this is particularly evident in the complex variety of this genre (where the average parse tree depth is 8.71), we can observe that also the simple one (i.e. Wikipedia articles) shows values that are much more higher than those reported by the simple varieties of all other genres. However, if scientific prose appears as the most difficult genre when sentence length and parse tree height are taken into account, it presents e.g. less complex verbal predicates in terms of number of dependents for verbal head (*avg_dep_verb*). With respect to this feature, Educational texts are those with the richest predicate structures, both in the easy and complex variety (2.16 and 2.25 respectively).

5.2 Studying the internal structure of text

We now focus on discussing the results of linguistic profiling carried out for the distinct sections of corpora, which were extracted according to the procedure described in Section 3. More specifically, to characterize the degree of variation within the different parts, and to evaluate how this variance derives from either genre-related or complexity-related features, we calculated two different statistical tests: *i)* the Wilcoxon

Table 2
An extract of linguistic features used for linguistic profiling across genres and complexity levels. For each feature it is reported the average value and (standard deviation) in the whole corpus (g), in the 'simple' (s) and in the 'complex' subcorpora (c). Values marked with exponent ^{n} mean that the corresponding feature is more stable for that corpus according to the coefficient of variation. The exponent index indicates the rank that these features have in the top ten positions of the ranking derived by the coefficient of variation.

features	Journalism			Scientific Prose		
	g	s	c	g	s	c
Raw text features						
sent_length	22.75 (11.39)	18.98 (7.65)	26.51 (13.13)	27.75 (13.63)	26.29 (13.96)	32.05 (11.62)
word_length	4.61^{2} (0.5)	4.58^{2} (0.52)	4.64^{2} (0.47)	5.07^{1} (0.62)	5.05^{1} (0.64)	5.13^{1} (0.57)
Morpho-syntactic features						
cpos_ADJ	6.21 (4.59)	6.1 (4.77)	6.32 (4.4)	8.63 (5.07)	8.58 (5.59)	8.8 (3.09)
cpos_ADV	4.47 (3.87)	3.75 (4.05)	5.19 (3.55)	3.55 (2.86)	3.57 (3.14)	3.49 (1.8)
cpos_CONJ	3.49 (2.93)	3.52 (3.32)	3.46 (2.49)	3.21 (2.32)	3.22 (2.51)	3.19 (1.63)
cpos_PREP	15.17 (5.56)	14.83 (5.94)	15.5 (5.13)	15.2^{9} (5.63)	14.64^{9} (5.92)	16.06 (4.52)
pos_PROPN	5.44 (6.12)	5.49 (6.39)	5.39 (5.19)	3.99 (8.07)	4.48 (9.04)	2.54 (3.59)
pos_NOUN	22.53^{7} (6.3)	23.51^{5} (6.32)	21.56^{7} (6.14)	26.04^{4} (7.3)	26.07^{5} (7.9)	25.94^{2} (4.93)
pos_VERB	10.93 (4.52)	11.29 (4.8)	10.57 (4.18)	8.08 (3.76)	7.98 (3.96)	8.37 (3.06)
Syntactic features						
dep_dobj	4.18 (3.33)	4.91 (3.87)	3.45 (2.48)	2.43 (1.94)	2.38 (2.14)	2.58 (1.2)
dep_subj	5.19 (3.15)	5.96 (3.53)	4.41 (2.51)	3.51 (2.22)	3.61 (2.44)	3.22 (1.86)
dep_mod	16.7 (6.63)	16.51 (7.35)	16.9 (5.81)	18.81^{8} (7.15)	18.35^{9} (7.62)	20.18^{7} (5.29)
avg_dep_noun	1.19^{5} (0.32)	1.19^{8} (0.33)	1.19^{5} (0.3)	1.2^{5} (0.34)	1.19^{4} (0.35)	1.24^{5} (0.3)
avg_dep_verb	2.02^{8} (0.61)	2.1^{7} (0.58)	1.93^{8} (0.62)	1.79^{10} (0.71)	1.78^{10} (0.75)	1.81 (0.61)
avg_dep_all	0.92^{1} (0.05)	0.92^{1} (0.03)	0.91^{1} (0.07)	0.85^{2} (0.18)	0.85^{2} (0.18)	0.86^{4} (0.19)
avg_links_l	2.07^{4} (0.49)	2.01^{4} (0.49)	2.13^{4} (0.47)	2.04^{6} (0.67)	2.03^{7} (0.71)	2.1^{6} (0.52)
parse_depth	6.87^{10} (2.24)	6.26^{6} (1.71)	7.5^{10} (2.51)	7.83^{7} (2.6)	7.56^{6} (2.6)	8.71^{8} (2.41)
parse_width	4.95^{9} (1.58)	4.84^{10} (1.45)	5.08^{9} (1.68)	5.27^{9} (2.02)	5.18^{8} (2.14)	5.56 (1.57)
link_pre	36.39^{6} (10.04)	37.51^{9} (10.46)	35.29^{6} (9.48)	26.77 (11.29)	27.12 (12.35)	25.74^{9} (7.19)
link_post	62.82^{3} (10.17)	62.2^{3} (10.49)	63.44^{3} (9.8)	65.95^{3} (16.09)	65.92^{3} (17.01)	66.06^{3} (12.99)
subord_depth	2.86 (2.22)	2.33 (2.13)	3.41 (2.18)	3.4 (2.14)	3.21 (2.25)	4.03 (1.63)
subord_width	1.59 (1.29)	1.3 (1.28)	1.89 (1.23)	1.86 (1.24)	1.79 (1.33)	2.11 (0.9)
sub_main	47.29 (32.76)	45.5 (37.25)	49.5 (27.29)	47 (27.86)	45.89 (30.23)	50.47 (18.61)

features	Narrative			Educational		
	g	s	c	g	s	c
Raw text features						
sent_length	18.99 (7.69)	17.62 (6.2)	20.4 (8.92)	28.02 (13.25)	22.79 (10.52)	31.96 (13.73)
word_length	4.27[1] (0.43)	4.22[1] (0.42)	4.33[2] (0.44)	4.55[2] (0.53)	4.33[2] (0.5)	4.72[2] (0.5)
Morpho-syntactic features						
cpos_ADJ	6.2 (4.7)	6.1 (4.76)	6.29 (4.66)	8.01 (4.29)	6.83 (4.08)	8.91 (4.24)
cpos_ADV	6.52 (4.23)	6.5 (4.26)	6.54 (4.21)	6.04 (3.8)	6.03 (4.21)	6.05 (3.47)
cpos_CONJ	5.24 (2.97)	5.31 (2.96)	5.16 (2.98)	4.56 (2.61)	4.17 (2.76)	4.86 (2.45)
cpos_PREP	12.06 (4.86)	11.86 (4.96)	12.28 (4.74)	14.21 (4.7)	13.36 (5.3)	14.85[9] (4.1)
pos_PROPN	4.99 (4.99)	5.28 (5.13)	4.7 (4.84)	1.78 (2.76)	2.2 (3.11)	1.46 (2.41)
pos_NOUN	17.93[9] (5.93)	18.07[9] (5.93)	17.78[9] (5.94)	21.33[5] (5.21)	20.85[7] (6.1)	21.69[4] (4.4)
pos_VERB	14.37[10] (4.89)	14.5[10] (4.8)	14.24 (4.99)	12.11[9] (3.89)	13.18[9] (4.26)	11.31 (3.37)
Syntactic features						
dep_dobj	4.53 (2.95)	4.66 (3.07)	4.39 (2.83)	3.78 (2.6)	3.93 (3.1)	3.67 (2.16)
dep_subj	6.17 (3.22)	6.42 (3.26)	5.92 (3.17)	5.24 (2.66)	5.64 (2.9)	4.94 (2.42)
dep_mod	15.05 (6.23)	14.76 (6.11)	15.35 (6.34)	16.46 (5.6)	15.23 (6.18)	17.4[10] (4.94)
avg_dep_noun	1.09 (0.38)	1.06 (0.36)	1.13[10] (0.39)	1.29[8] (0.37)	1.2[10] (0.43)	1.35[6] (0.3)
avg_dep_verb	2.1[5] (0.59)	2.12[6] (0.58)	2.09[5] (0.6)	2.21[6] (0.54)	2.16[5] (0.56)	2.25[7] (0.52)
avg_dep_all	0.89[2] (0.09)	0.89[2] (0.09)	0.9[1] (0.08)	0.93[1] (0.05)	0.91[1] (0.05)	0.94[1] (0.05)
avg_links_l	2.01[4] (0.48)	1.97[4] (0.47)	2.06[4] (0.49)	2.22[4] (0.53)	2.08[6] (0.55)	2.33[5] (0.48)
parse_depth	6.28[7] (1.86)	6.02[5] (1.64)	6.54[10] (2.03)	7.61[8] 2.56	6.44[8] (1.89)	8.45 (2.7)
parse_width	4.5[8] (1.37)	4.33[7] (1.22)	4.68[8] (1.49)	5.62 (2.02)	4.98 (1.89)	6.08 (2.02)
link_pre	38.65[6] (11.43)	38.92[8] (11.23)	38.39[6] (11.63)	36.56[7] (9.24)	39.27[4] (9.75)	34.52[8] (8.27)
link_post	59.81[3] (12.11)	59.29[3] (12.04)	60.33[3] (12.17)	62.52[3] (9.53)	59.5[3] (10.14)	64.8[3] (8.36)
subord_depth	2.79 (1.71)	2.56 (1.64)	3.03 (1.75)	3.57 (2.19)	2.81 (1.88)	4.2 (2.24)
subord_width	1.56 (1.05)	1.41 (1)	1.7 (1.09)	2.15 (1.4)	1.64 (1.22)	2.56 (1.41)
sub_main	50.76 (28.92)	48.68 (29.41)	52.89 (28.31)	54.53 (28.6)	52.07 (31.87)	56.43 (25.22)
sub_minor	8.92 (13.5)	7 (11.1)	1.12 (7.37)	12.74 (15.52)	8.94 (13.78)	16.27 (16.6)

rank-sum test and *ii)* the Spearman's correlation test. As previously stated, each metric was calculated between the average value of features extracted from two consecutive sections and from the first and the last section of each document. Results of this comparison are reported in Table 3, which displays all features that 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.

As a first remark, we clearly notice that a 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 texts of the Educational domain, which is probably due to the heterogeneous nature of this 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. For instance, it is characterized by shorter sentences (Figure 2), likely due to the presence of the title in both newspaper and scientific articles, and by a distinctive distribution of Parts–of–speech (Figure 3). With this respect, this data is consistent with other studies in the literature, e.g. (Voghera 2005), and also with what we observed in the global analysis of our corpora reported in paragraph 5.1, 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 syntactic structure 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 *la Repubblica*. Specifically, the first section contains less embedded sentences (*parse_depth*: 1st section: 5.55; 2nd section: 7.76), and a lower presence of subordinate clauses, which appear as structurally simpler e.g. in terms of depth (*subord_depth*: 1st section: 1.67; 2nd section: 3.5) and width (*subord_width*: 1st section: 0.94; 2nd section: 1.97). Conversely, for the simple variant of this genre (i.e. the articles of the easy-to-read newspaper *Due Parole*), 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 (as reported in Table 2, the distribution of subordinate clauses with respect to the main clause (*sub_main*) is 45.5.)

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 the hierarchical structure of the parse tree (e.g. *parse_width, avg_dep_all*) and to the presence and level of embedding of subordinate clauses (*subord_depth, subord_width*). According to these data, we can conclude that the journalistic genre has a more rigorous structure and that it is possible to capture the

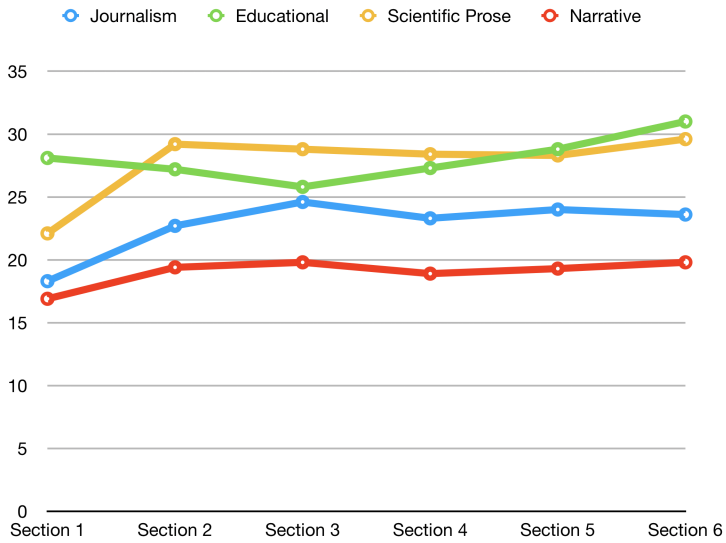

Figure 2
Average sentence length in the 6 sections across genres.

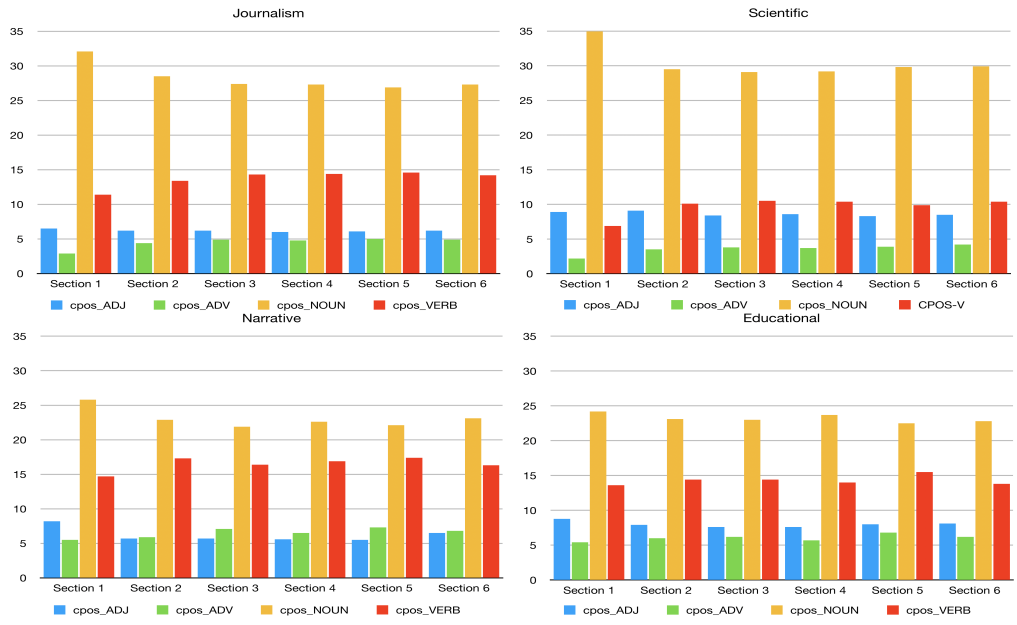

Figure 3
Distribution of lexical parts-of-speech in the four genres.

boundaries between different parts by using linguistic features that are not related to the content of the article.

Table 3
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.

features		Journalism						Scientific Prose					
		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													
sent_length	g	✓✓	✓*	-*	-*	-*	✓✓	✓✓	-	-	-*	-*	✓✓
	s	✓✓	-	-	-	-	-	✓✓	-	-	-	-	✓✓
	c	✓✓	-	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*
word_length	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_PROPN	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	✓✓	-	✓	-	-	✓✓	-*	-*	-*	-*	-*	-
parse_depth	g	✓✓	-*	-*	-*	-*	✓✓	✓✓	-	-	-	✓*	✓✓
	s	-	-	-	✓	-	✓	✓✓	-	-	✓	-*	✓✓
	c	✓✓	-	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*
parse_width	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓
	s	✓✓	-	-	-	-	-	✓✓	-	-	-	-	✓✓
	c	✓✓	-	✓	-*	-	✓✓	-*	-*	-*	-*	-*	-*
avg_dep_all	g	✓✓	✓	-*	-*	-*	✓✓	✓✓	-*	-*	-*	✓*	✓✓
	s	✓	-	-	-	-*	-	✓✓	-	-	-	-	✓✓
	c	✓✓	✓	-*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*
Subordination features													
subord_depth	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	✓	✓✓
	s	-	-	-	-	-	✓	✓✓	-	-	-	-	✓✓
	c	✓✓	-	-	-	-	✓✓	-*	-*	-*	-*	-*	-*
subord_width	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	✓	✓✓
	s	-	-	-	-	-	✓	✓✓	-	-	-	-	✓✓
	c	✓✓	-	-	-	-	✓✓	-*	-*	-	-*	-*	-*
sub_main	g	✓✓	✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓
	s	-	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓
	c	✓✓	-	-	-	-	✓✓	-*	-*	-*	-*	-*	-*
sub_minor	g	✓✓	-*	-	-*	-	✓✓	-	-	-	-	-	✓
	s	-	-	-	-	-	✓	-	-	-	-	-	✓
	c	✓✓	-	-	-	-	✓✓	-*	-*	-	-*	-	-

features		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
Raw text features													
sent_length	g	-	-	-	-	-	✓*	-*	-*	-*	-*	-*	-*
	s	-*	-	-	-	-	✓	-*	-*	-*	-*	✓*	-*
	c	-	-	-	-	-	✓	-*	-	-*	-	-	-
word_length	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_PROPN	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_dep_all	g	✓	-*	-*	-	-*	✓	-*	-*	-*	-*	-*	-*
	s	-	-	-*	-	-*	✓	-	-*	-*	-*	-*	-*
	c	-	-*	-*	-	-*	✓	-*	-*	✓	-*	-*	-*
Subordination features													
subord_depth	g	-*	-	-*	-*	-*	-	-	-	-*	-*	-	-
	s	-	-	-	-*	-*	✓✓	-	-	-*	-*	-*	✓
	c	-	-	-	-*	-	-	-	✓	-	-	-	-
subord_width	g	-	-	-	-	-	-	-	-	-*	-	-	-
	s	-	✓	-	-*	-*	✓	-	✓	-*	-*	-*	-
	c	-	-	-	-*	-	-	-	-	-	-	-*	-
sub_main	g	-	-	-	-	-*	✓✓	-	-	-	-	-	-
	s	-	-	-	-	-*	✓✓	-	✓*	✓✓*	-*	✓*	-
	c	-	-	-	-	-	-	-	✓	-	-	-	-
sub_minor	g	-	-	-	-	✓	-	-	-*	✓	-	-	✓
	s	-	-	-	-	-*	-*	-	-	✓	-	-*	-*
	c	✓	-	-	-*	-	-	-	-	-	-*	-	✓

6. 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 comparative 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 representative of a given genre and language variety. From a linguistic perspective, this suggests that the study of genre and register variation can benefit by inspecting corpora from this fine-grained perspective. In this respect, we intend to carry out further analyses to prove the validity of this approach and how scalable it is when applied to novel texts of the same genres here considered. A first step in this direction is surely to assess the accuracy of our distinction into six parts, which was mainly driven by corpus-based considerations in terms of average document length. In particular, we would like to compare whether the findings obtained applying this splitting methodology are in line with those deriving form a manually-based annotation, in which the sections are identified according to their structural coherence as perceived by readers.

We believe that a better understanding and computational modeling of linguistic phenomena characterizing the introductory, middle and conclusive parts of a text can also serve to enhance automatic genre classification, as well as a number of NLP-based applications devoted to modeling style. For instance, in the educational domain, it could be used as a part of intelligent tutoring systems able to provide detailed feedback to students in writing courses and to support the development of automatic self-assessment tools. Also in the field of Natural Language Generation, such a rich feature-based description of the internal profile of texts can be useful to inform the automatic generation of texts, which are not only semantically meaningful and coherent, but also compliant with the stylistic fingerprints of a specific genre and level of complexity.

References

Argamon, Shlomo E. 2019. Computational register analysis and synthesis. *Register Studies*, 1(1):100–135.

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

Barbagli, Alessia, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2015. Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati. *Italian Journal of Computational Linguistics (IJCoL)*, 1(1):99–117.

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

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: investigating language structure and use.* Cambridge University Press, Cambridge.

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

Brunato, Dominique, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods of Natural Language Processing*, Brussels, Belgium, October-November.

Cimino, Andrea, Martijn Wieling, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Identifying predictive features for textual genre classification: the key role of syntax. In *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, Rome, 11-13 December 2017.

Collins-Thompson, Kevyn. 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):97–135.

Crossley, Scott A., Kyle B. Dempsey, and Danielle S. McNamara. 2011. Classifying paragraph types using linguistic features: Is paragraph positioning important? *Journal of Writing Research*.

Daelemans, Walter. 2013. Explanation in computational stylometry. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pages 451–462. Springer Berlin Heidelberg, march.

Dell'Orletta, Felice. 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.

Finn, Aidan and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the Association for Information Science and Technology*, 57:1506–1518.

Gildea, Daniel. 2001. Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA, USA, June.

Lu, Xiaofei. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Lubetich, Shannon and Kenji Sagae. 2014. Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160, Dublin, Ireland, August.

Malmasi, Shervin, Evanini Keelan, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP, held at EMNLP 2017*, Copenhagen, Denmark, September.

Montemagni, Simonetta. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, XLII(1):145–172.

Montemagni, Simonetta, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessando Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rofolfo Delmonte. 2003. Building the italian syntactic-semantic treebank. In A. Abeille, editor, *Building and Using syntactically annotated corpora*.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

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

Prud'hommeaux, Emily T., Brian Roark, Lois M. Black, and Jan van Santen. 2011. Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, held at ACL HLT*, pages 88–96, Portland, Oregon, USA, June.

Roark, Brian, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, held at ACL 2007*, pages 1–8, Prague, Czech Republic, June.

Rouhizadeh, Masoud, Richard Sproat, and Jan van Santen. 2015. Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics*, pages 117–123, Denver, Colorado, June.

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

van Halteren, Hans. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics (ACL04)*, pages 200–207, Barcelona, Spain, July.

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