Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach

Giulia Venturi [*]

Abstract. *In this paper, the author carried out the linguistic profiling of a corpus of different types of Italian legal texts exemplifying different sub-varieties of Italian legal language by relying on a wide range of different linguistic features (lexical, morpho-syntactic and syntactic) automatically extracted from the output of a multi-level automatic linguistic analysis of texts. The devised comparative approach allowed investigating the linguistic variation* i) *between the considered corpus of legal texts and a corpus of newspaper articles representative of Italian ordinary language and* ii) *among the considered types of legal texts (legislative acts, administrative acts, the Italian Constitution and legal cases). Achieved results can provide the starting point to identify areas of lexical, morpho-syntactic and/or syntactic complexity within a legal text in order to assess its readability as well to perform a number of different computational forensic linguistics tasks.*

Keywords: Linguistic profiling, Natural Language Processing, legal language analysis, legal genres, readability assessment, syntactic complexity.

Introduction

Over the last few years, the use of Natural Language Processing (NLP) tools and techniques has spread within computational forensic linguistics studies. In spite of the fact that they address different purposes, such as authorship attribution (Sousa-Silva *et al.*, 2010) or automatic deception detection (Fornaciari and Poesio, 2013), these studies share a common approach: they succeed in their specific goal by exploiting the distribution of a number of different linguistic characteristics automatically extracted from the linguistically analysed text. More generally, this is the basis of the so-called "linguistic profiling" within which "the occurrences in a text of a large number of linguistic features, either individual items or combinations of items, are counted" (van Halteren, 2004: 202).

[*] Instituto di Linguistica Computazionale "Antonio Zampolli", CNR

In this paper, the *linguistic profiling* of a corpus of different types of Italian legal texts exemplifying different sub-varieties of Italian legal language is carried out by relying on NLP-enabled linguistic features automatically extracted from text. It stems from studies on register analysis, such as those carried out by Biber and colleagues (1993; 1998; 2009), and in particular from studies focussed on the linguistic characteristics specific to different types of legal texts, i.e. legal texts pursuing different communicative purposes (Bathia, 1993). Unlike previous studies, feature extraction was carried out on the output of a multi-level automatic linguistic analysis: a wide typology of selected features was looked for within each level of the automatic linguistic analysis (lexical, morpho-syntactic and syntactic). The eventual goal is to demonstrate how automatic linguistic analysis can provide a useful starting point for the linguistic profiling of legal texts.

The author followed a comparative approach with two specific goals. The first goal consists in finding significative linguistic variation between the corpus of legal texts under consideration and a corpus of newspaper articles representative of Italian ordinary language. This is quite crucial, since legal language *differs* from ordinary language, but it is not dramatically *independent* from every day speech (Mortara Garavelli, 2001; Rovere, 2005). The second goal is the investigation of the linguistic characteristics that make various types of legal texts different and distinguishable. The interest in pursuing this goal relies on the widely acknowledged fact that the "term 'language of the law' encompasses several usefully distinguishable genres" and that "these genre distinctions are also reflected in the lexico-grammatical, semantico-pragmatic, and discoursal resources that are typically and conventionally employed to achieve successful communication in various legal settings" (Bathia, 1987: 227). For this purpose, we compared the different types of texts included in the corpus of legal texts described here.

Methodology

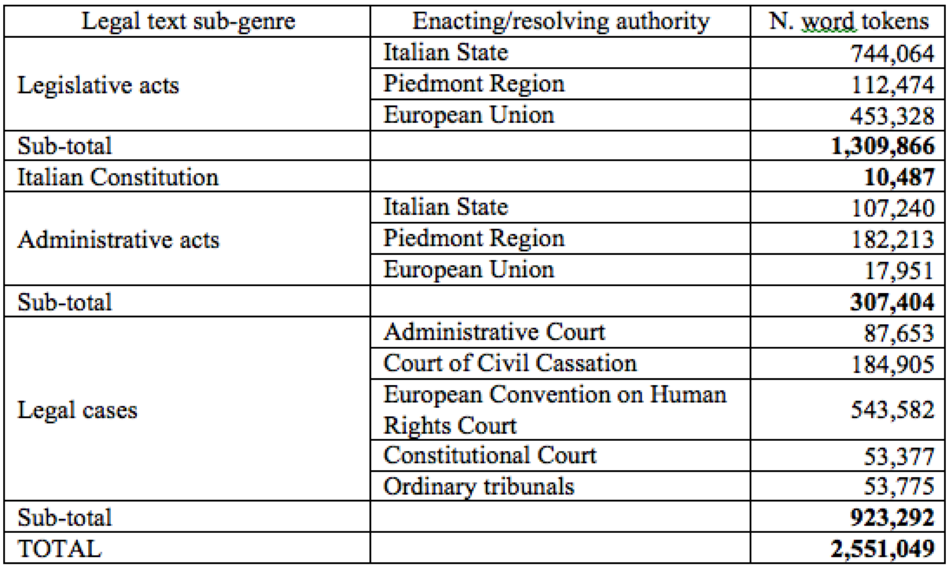
The approach to linguistic profiling devised here stems from Dell'Orletta *et al.* (2013) and includes three main ingredients:

- a collection of legal texts exemplifying different sub-varieties of Italian legal language and of texts representative of Italian general language used as reference corpus,
- a collection of NLP tools performing the automatic linguistic analysis of texts,
- a method of linguistic profiling which allows the comparison of different corpora according to the distribution of a set of linguistic features.

Corpora

For the specific concerns of this study, we built a corpus of legal texts exemplifying different sub-varieties of Italian legal language, i.e. legislative acts, administrative acts, the Italian Constitution and legal cases. Following the legal text classification suggested by Bathia (1987), they belong to two main classes of documents used in two different legal settings. While the legislative and administrative acts, as well as the Italian Constitution, are types of documents used in a "legislative setting", the legal cases are typically used in a "juridical setting". According to the classification of the Italian legal texts proposed by Mortara Garavelli (2001: 19–34), the Italian Constitution belongs to the class of "normative" texts (used in a "legislative setting"). However, it was analysed here separately from the legislative texts in order to highlight its linguistic specificity with respect to other normative texts.

Table 1. The corpora of legal texts.



Legal text sub-genre	Enacting/resolving authority	N. word tokens
Legislative acts	Italian State	744,064
	Piedmont Region	112,474
	European Union	453,328
Sub-total		**1,309,866**
Italian Constitution		**10,487**
Administrative acts	Italian State	107,240
	Piedmont Region	182,213
	European Union	17,951
Sub-total		**307,404**
Legal cases	Administrative Court	87,653
	Court of Civil Cassation	184,905
	European Convention on Human Rights Court	543,582
	Constitutional Court	53,377
	Ordinary tribunals	53,775
Sub-total		**923,292**
TOTAL		**2,551,049**

Table 1 shows the internal partition of this Italian legal text corpus. As can be seen, both legislative and administrative corpora include acts enacted by the Italian State, the Piedmont region and the European Union. The two corpora are respectively made up of legislative acts, such as national and regional laws, European directives, legislative decrees, and administrative acts, such as ministerial circulars and decisions. The collection of legal cases contains texts resolved by various Italian administrative courts, the Italian Constitutional court, the Italian Court of Civil Cassation, the European Convention on Human Rights Court, and by various Italian ordinary tribunals. They all concern the principle of state *liability*, i.e. the general principle established by the European Court of Justice according to which Member States should pay compensation to individuals who suffered a loss by reason of the State failing to comply with the European law (Lazari, 2005). Finally, the Italian Constitution was analysed in its 1947 original version.

According to the comparative approach devised here, we chose a corpus of newspaper articles taken as representative of general Italian language as a baseline for comparison. The choice of the journalistic prose as reference genre is inspired by the work of Rovere (2005) who investigated the morpho-syntactic and syntactic differences of a corpus of different types of Italian legal texts against a corpus of Italian newspapers, focussing on the relationship between the semantic and syntactic valencies of some verbs.

However, unlike the case of Rovere, two newswire corpora were considered here: a collection of articles taken from “La Repubblica” daily newspaper and from “Due Parole”[1], a newspaper written by Italian linguistic experts in text simplification using a controlled language (Piemontese, 1996). According to their linguistic peculiarities, as empirically demonstrated by Dell’Orletta and Montemagni (2012) and Dell’Orletta *et al.* (2011b), the two corpora can be seen as two opposite poles of the same textual genre: “Due Parole” represents a newspaper explicitly written using plain language while “La Repubblica” articles represent the opposite extreme being written using everyday language. They were both taken as reference corpora here since they allow extensive investigation of the linguistic peculiarities of legal texts.

Natural Language Processing tools

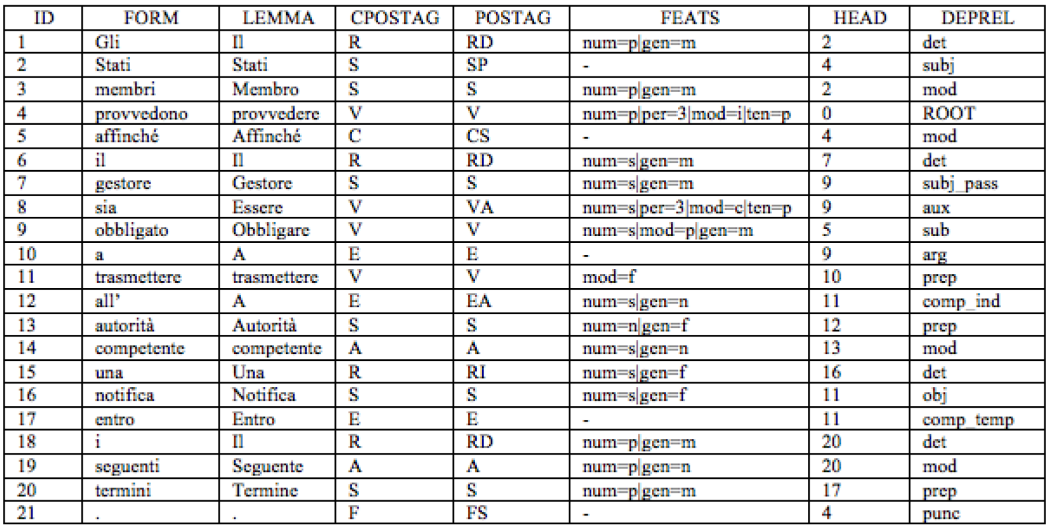
All the corpora were automatically analysed by a collection of statistical NLP tools jointly developed by the Institute of Computational Linguistics "Antonio Zampolli" in Pisa (ILC-CNR) and the University of Pisa. They were morpho-syntactically tagged by the Part-Of-Speech tagger described in Dell'Orletta (2009) and syntactically annotated by the DeSR parser, the dependency parser described in Attardi (2006).

The tools are able to make evident the implicit linguistic information contained in texts by *annotating* them at increasingly complex levels of analysis. Namely, they split the whole text into sentences, segment each sentence into orthographic units (tokens), assign all possible morphological analyses to each token, assign the appropriate morpho-syntactic interpretation in the specific context and identify existing syntactic dependency relations between tokens (e.g. subject, object, etc.). For example, the following sentence is annotated as seen in Table 2[2]:

(1) Gli Stati membri provvedono affinché il gestore sia obbligato a trasmettere all'autorità competente una notifica entro i seguenti termini. ('Member States shall require the operator to send the competent authority a notification within the following time-limits'.)

In Table 2, it can be noted that each word form (in the column headed FORM), univocally marked by a numerical identifier (column ID), is associated with its corresponding lemma (column LEMMA), its coarse- (column CPOSTAG) and fine-grained (column POSTAG) part-of-speech and its morphological treats (column FEATS). Moreover, the annotation makes explicit the head of the dependency syntactic relation in which each word is involved (column HEAD) and the type of dependency relation (column DEPREL). For example, Table 2 shows that the word *notifica* ('notification') is the object (obj) of the verb *trasmettere* ('send').

Table 2. Example of an annotated sentence in CoNLL format.



ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Gli	Il	R	RD	num=p|gen=m	2	det
2	Stati	Stati	S	SP	-	4	subj
3	membri	Membro	S	S	num=p|gen=m	2	mod
4	provvedono	provvedere	V	V	num=p|per=3|mod=i|ten=p	0	ROOT
5	affinché	Affinché	C	CS	-	4	mod
6	il	Il	R	RD	num=s|gen=m	7	det
7	gestore	Gestore	S	S	num=s|gen=m	9	subj_pass
8	sia	Essere	V	VA	num=s|per=3|mod=c|ten=p	9	aux
9	obbligato	Obbligare	V	V	num=s|mod=p|gen=m	5	sub
10	a	A	E	E	-	9	arg
11	trasmettere	trasmettere	V	V	mod=f	10	prep
12	all'	A	E	EA	num=s|gen=n	11	comp_ind
13	autorità	Autorità	S	S	num=n|gen=f	12	prep
14	competente	competente	A	A	num=s|gen=n	13	mod
15	una	Una	R	RI	num=s|gen=f	16	det
16	notifica	Notifica	S	S	num=s|gen=f	11	obj
17	entro	Entro	E	E	-	11	comp_temp
18	i	Il	R	RD	num=p|gen=m	20	det
19	seguenti	Seguente	A	A	num=p|gen=n	20	mod
20	termini	Termine	S	S	num=p|gen=m	17	prep
21	.	.	F	FS	-	4	punc

It should be mentioned here that even if the NLP tools exploited in this study represent the state of the art for the Italian language (see the results of the Evalita 2009 evaluation campaign[3], however their performances are affected by the particular type of texts at hand. Since (Gildea, 2001), it is widely acknowledged that state-of-the-art linguistic annotation tools suffer from a dramatic drop of accuracy when tested on domains outside of the data from which they were trained or developed. As recently testified by the "Domain Adaptation Track" (Dell'Orletta *et al.*, 2012a) and by "The SPLeT-2012 Shared Task on Dependency Parsing of Legal Text" (Dell'Orletta *et al.*, 2012b), the legal domain does not represent an exception. In order to cope with this problem, for the specific concerns of this study, the statistical NLP tools were specialised by combining two training sets: the ISST-TANL treebank consisting of newspaper articles (Dell'Orletta *et al.*, 2012b) taken as representative of general language usage, and the TEMIS corpus (Venturi, 2012), a syntactically annotated corpus of Italian legislative and administrative texts. This allowed maintaining the state-of-the-art performance of the NLP tools exploited here.

Comparative approach and linguistic features

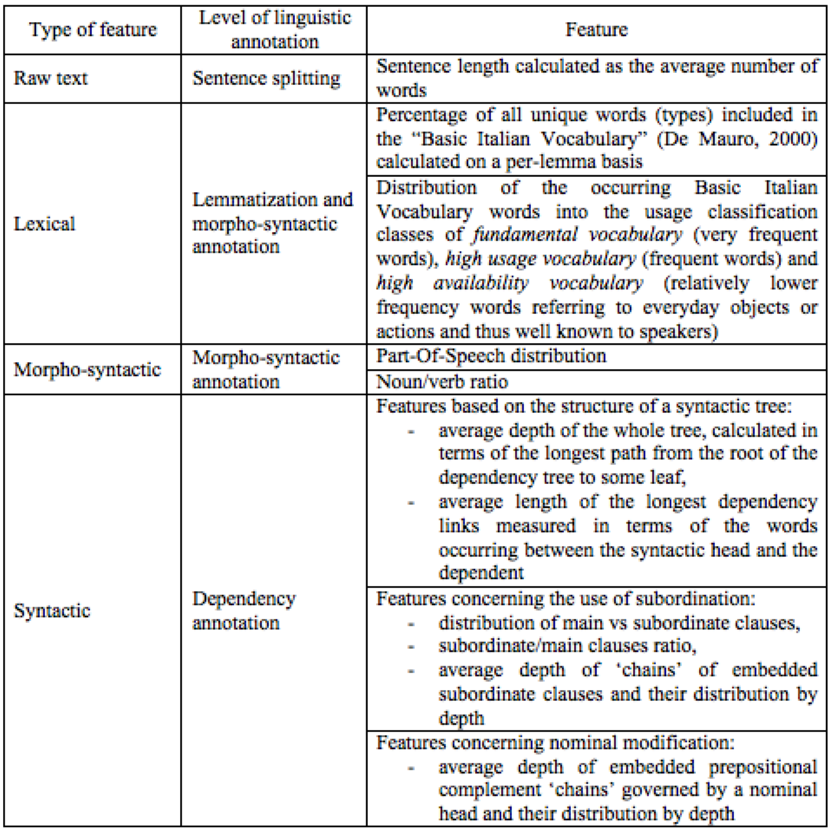
The comparative approach followed in this study stems from the literature on register variation and is mostly inspired by the work of Biber who claims that "we need a baseline for comparison to know whether the use of a linguistic feature in a register is rare or common" (Biber *et al.*, 1998). For the specific concerns of this study, the corpora illustrated in Section have been compared at two different levels. Firstly, the distribution of some linguistic features was comparatively observed within the whole collection of legal texts and the newswire corpora. This perspective of analysis has allowed the highlighting of how and to what extent legal language differs from ordinary language. Secondly, the different types of legal texts were compared at different levels of specificity in order to investigate the linguistic variation between documents used in a "legislative" and "juridical setting" and between documents enacted or resolved by different authorities. This second level of analysis allowed us to show the *multiform and complex* nature specific to the legal language (Cortelazzo, 1997) highlighting the peculiarities of different legal language sub-varieties.

Starting from the output of the automatic linguistic analysis, all the corpora were searched for with respect to four classes of linguistic features: raw text, lexical, morpho-syntactic and syntactic. As illustrated in 3, this four-fold partition closely follows the different levels of linguistic annotation and it was prompted by Biber's idea that "linguistic features from all levels function together as underlying dimensions of variation and [. . .] there are systematic and important linguistic differences among registers with respect to these dimensions" (Biber, 1993: 220–221).

Two different criteria have guided the choice of these linguistic features. Firstly, some of them are contained in the Italian "Guide to drafting administrative acts" devoted to suggesting how to draft acts in a plain language[4]. Gathering the suggestions put by the "Directive on the simplification of the language of administrative acts" of the Ministry for Civil Service Reform and by the "Rules and suggestion to drafting legislative acts" adopted by the Italian local authorities, the "Guide" is today the most up-to-date collection describing the lexical, morpho-syntactic and syntactic characteristics that a legal document is expected to conform to in order to be written in a *plain, simple and comprehensible* language.

Secondly, this work would like to follow the methodology devised by Dell'Orletta and Montemagni (2012) who demonstrated the high discriminative power of the set of linguistic features considered here to monitor diastratic, diamesic and diaphasic varieties of Italian language. Their results were also confirmed when these features were adopted to linguistically profile Italian educational materials, such as textbooks for primary and

Table 3. The considered linguistic features.



Type of feature	Level of linguistic annotation	Feature
Raw text	Sentence splitting	Sentence length calculated as the average number of words
Lexical	Lemmatization and morpho-syntactic annotation	Percentage of all unique words (types) included in the "Basic Italian Vocabulary" (De Mauro, 2000) calculated on a per-lemma basis
		Distribution of the occurring Basic Italian Vocabulary words into the usage classification classes of *fundamental vocabulary* (very frequent words), *high usage vocabulary* (frequent words) and *high availability vocabulary* (relatively lower frequency words referring to everyday objects or actions and thus well known to speakers)
Morpho-syntactic	Morpho-syntactic annotation	Part-Of-Speech distribution
		Noun/verb ratio
Syntactic	Dependency annotation	Features based on the structure of a syntactic tree: - average depth of the whole tree, calculated in terms of the longest path from the root of the dependency tree to some leaf, - average length of the longest dependency links measured in terms of the words occurring between the syntactic head and the dependent
		Features concerning the use of subordination: - distribution of main vs subordinate clauses, - subordinate/main clauses ratio, - average depth of 'chains' of embedded subordinate clauses and their distribution by depth
		Features concerning nominal modification: - average depth of embedded prepositional complement 'chains' governed by a nominal head and their distribution by depth

high school and written productions of learners of Italian (Dell'Orletta *et al.*, 2011a), and to assess the readability of newspaper texts (Dell'Orletta *et al.*, 2011b) and of different text genres (Dell'Orletta *et al.*, 2012c).

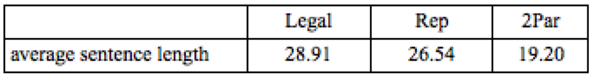
Legal texts vs newspaper articles: linguistic peculiarities

The linguistic characteristics specific to the corpus of legal texts taken as a whole are reported and discussed in what follows. They provide first insights into the peculiarities of legal language compared to ordinary language.

Raw and lexical text features

According to one of the first suggestions contained in the "Guide to drafting administrative acts", a legal text should contain *short* sentences. As Table 4 shows, the comparison of the legal and newswire corpora revealed that on the contrary the whole legal corpus (*Legal* in the Table) contains sentences longer than those occurring both in "La Repubblica" (*Rep*) and in "Due Parole" (*2Par*).

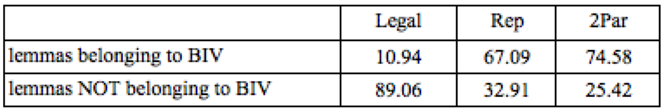
Table 4. Average sentence length in legal and newswire corpora.



	Legal	Rep	2Par
average sentence length	28.91	26.54	19.20

A similar recommendation towards a text written in a plain language holds for the lexical profile that a legal text should exhibit. According to the "Guide", a legal text should mainly contain words belonging to the “Basic Italian Vocabulary” since they are more frequently used and consequently more comprehensible. As Table 5 reports, the legal corpus contains a rather lower percentage of “Basic Italian Vocabulary” (BIV) lemmas (calculated in terms of types, as reported above in Table 3) with respect to newswire corpora.

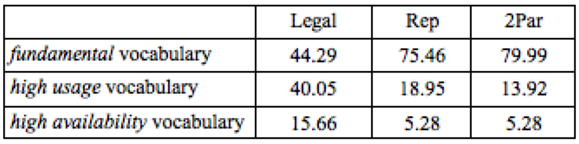
Table 5. % of lemmas (types) belonging to the “Basic Italian Vocabulary” in legal and newswire corpora.



	Legal	Rep	2Par
lemmas belonging to BIV	10.94	67.09	74.58
lemmas NOT belonging to BIV	89.06	32.91	25.42

When we further compare the corpora with respect to the percentage distribution of the occurring “Basic Italian Vocabulary” lemmas into the usage classification classes (i.e. *fundamental*, *high usage* and *high availability*), some interesting results can be observed. As Table 6 shows, the legal corpus not only contains a lower percentage of fundamental vocabulary, but the difference between the percentage of this class and the *high usage* and *high availability* vocabulary is lower. In the legal corpus the difference between the distribution of fundamental and high usage vocabulary is of 4.24 percentage points, while in “La Repubblica” corpus the same difference is of 56.51 percentage points and in “Due Parole” it is of 66.07 points. Such a difference shows the particular tendency in the legal corpus towards the use of *high usage* and *high availability* vocabulary.

Table 6. % of lemmas (types) distributed in the three usage classes in legal and newswire corpora.



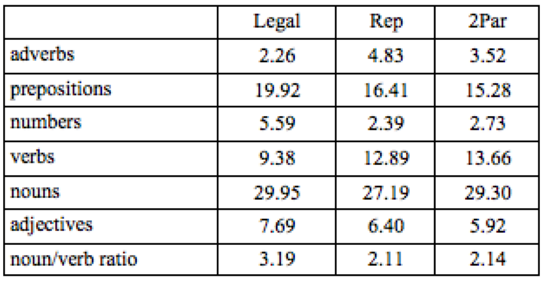
	Legal	Rep	2Par
fundamental vocabulary	44.29	75.46	79.99
high usage vocabulary	40.05	18.95	13.92
high availability vocabulary	15.66	5.28	5.28

Morpho-syntactic features

At the morpho-syntactic annotation level we observe a different distribution of Parts-of-Speech (PoS) categories occurring in the legal and newswire corpora. The percentage distribution of the considered PoS are reported in Table 7. It can be noted that the legal corpus has a higher percentage of **prepositions**, **numbers**, **nouns** and **adjectives** while it contains a significantly lower percentage of **verbs** and **adverbs**.

Interestingly, these results are in line with the literature on register variation. According to Biber (1993), both a high co-occurrence of **nouns**, **prepositions** and **adjectives** and different noun/verb ratio values represent two significant dimensions of variation among textual genres. Concerning this latter characteristic, Biber (1993) found that highly informative genres such as academic prose are characterised by a higher noun/verb ratio with respect to texts representative of fiction prose or with respect to speech. By relying on the output of a collection of NLP tools, a similar tendency was observed by Montemagni (2013) for the Italian language. She demonstrated that a corpus of fiction texts and of speech transcriptions have a lower noun/verb ratio than to a corpus of newspaper articles.

Table 7. Morpho-syntactic features in legal and newswire corpora.



	Legal	Rep	2Par
adverbs	2.26	4.83	3.52
prepositions	19.92	16.41	15.28
numbers	5.59	2.39	2.73
verbs	9.38	12.89	13.66
nouns	29.95	27.19	29.30
adjectives	7.69	6.40	5.92
noun/verb ratio	3.19	2.11	2.14

The different distribution of nouns, prepositions and adjectives in the legal and newswire corpora suggests that we are dealing with two different linguistic varieties. The higher noun/verb ratio provides further evidence that legal texts are even more informative than newspaper articles. It is a straight consequence of the higher percentage of **nouns** and, above all, of the quite low percentage occurrence of **verbs** occurring in the legal texts.

The significant higher occurrence of **numbers** in the legal texts is a further peculiarity of this genre. It is due to several different reasons, e.g. the ample use of numbers corresponding to textual partition numbering (e.g. article, paragraph), the occurrence of dates, the citations between legal cases or legislative acts also expressed through numerical identifiers, etc.

Syntactic features

As reported above in Table 3, in this study we considered three types of syntactic features concerning: the structure of a syntactic tree, the use of subordination and the nominal modification.

Features based on the structure of a syntactic tree

Two features have been taken into account: the average depth of the syntactic tree and the average length of the longest dependency links.

Consider the two following sample sentences extracted from the legal corpus:

(2) I proprietari, possessori o detentori a qualsiasi titolo dei beni indicati al comma 1, hanno l'obbligo di sottoporre alla Regione i progetti delle opere di qualunque genere che intendano eseguire, al fine di ottenere la preventiva autorizzazione. ('The owners, possessers or holders on whatever basis of the goods mentioned in paragraph 1, have the obligation to submit to the Region the projects of the works of any kinds they plan to carry out, in order to obtain prior authorization.')

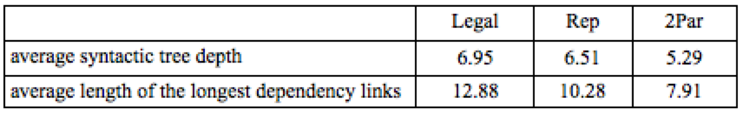
(3) Chiunque immette sul mercato i preparati pericolosi di cui al presente decreto, in violazione delle disposizioni in tema d'imballaggio e di etichettatura di cui agli articoli 8, 9 e 10, nonché in violazione delle disposizioni sulla classificazione di cui all'articolo 3, è punito con l'ammenda da euro centoquattro a euro cinquemilacentosessantacinque. ('Anyone who places on the market dangerous preparations provided for in this decree, in violation of the provisions on packaging and labelling referred to in articles 8, 9 and 10, and in violation of the provisions on the classification referred to in article 3, is punished by a fine of one hundred and four euro to euro five thousand one hundred sixty-five.')

In (2), the syntactic tree has a maximum depth=8. As Figure 1 shows, it is calculated as the sequence of eight consecutive dependency links[5], i.e. direct object (*obj*), argument (*arg*), preposition (*prep*), direct object (*obj*), complement (*comp*), preposition (*prep*) and relative modifier (*mod_rel*), which starts from the root of the syntactic tree *hanno* ('have') and ends to the leaf *eseguire* ('carry out').

In line with the literature on measuring dependency distance, the length of the longest dependency link is measured here in terms of "intervening words" (Hudson, 1995: 16) between a dependent and its parent. In (3), the longest dependency link is 47 tokens long[6]: it is the subject relation between the dependent *chiunque* ('anyone') and its governing head corresponding to the syntactic root of the sentence (*punito*, 'punished').

As can be seen in Table 8, legal sentences are characterised by deeper syntactic trees and much longer dependency links.

Table 8. Features based on the structure of syntactic trees in legal and newswire corpora.



	Legal	Rep	2Par
average syntactic tree depth	6.95	6.51	5.29
average length of the longest dependency links	12.88	10.28	7.91

These results suggest that within legal texts there is a more complex syntactic structure with respect to newspaper articles. Not only can a deep syntactic tree be indicative of increased sentence complexity as stated by e.g. Frazier (1985) and Gibson (1998), but the same holds for long dependency links. In a dependency representation scenario, Hudson (1995: 15-16) claims that the "dependency distance", i.e. "the distance between words and their parents, measured in terms of intervening words", "might be relevant to how hard a sentence is to process". Accordingly, the greater the dependency distance, the more complex is the sentence. This is in line with the findings of studies carried out in the cognitive and psycholinguistic field. In particular, it has been ascertained that "there is a finite span of immediate memory and that [...] this span is about seven items in length" (Miller, 1956: 9). It follows that it is perceptually costly to carry on analysing sentences with long dependencies.

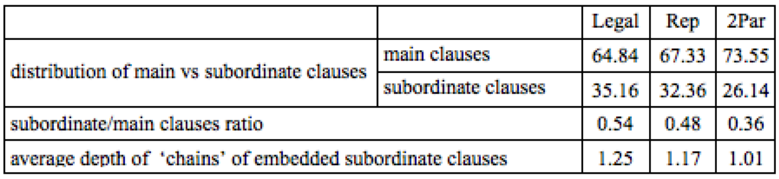
Features concerning the use of subordination

Concerning the use of subordination, the "Guide" recommends a limit on the subordinate clauses. This follows from the literature which relates syntactic complexity to the occurrence of embedding structures and in particular to the presence of subordinate clauses. According to Beaman (1984: 45) "syntactically complex authors [...] use longer sentences and more subordinate clauses". Typically, the use of parataxis is preferable to a hypotactic structure since a coordinated construction is in principle more easy-to-read and comprehensible than a subordinate one (Piemontese, 1996) which, on the contrary, is cognitively more complex (Givón, 1991).

In contrast with the recommendations of the "Guide", as Table 9 shows, the results of the whole legal corpus are characterised by a higher percentage of subordinate clauses organised in long 'chains', and consequently by a higher subordinate/main clauses ratio.

Moreover, the legal corpus contains a higher percentual distribution of deeply embedded subordinate clauses. For example, 'chains' of 3 embedded subordinate clauses constitute 3.51% of the total amount of 'chains' of subordinate clauses occurring in the whole legal corpus while they have a coverage of only 2.89% in "La Repubblica" corpus and of 1.32% in "Due Parole". This quite differs from the distributions found in "Due Parole", less sharp but still statistically significant when the "La Repubblica" ratio is considered.

Table 9. Use of subordination in legal and newswire corpora.



		Legal	Rep	2Par
distribution of main vs subordinate clauses	main clauses	64.84	67.33	73.55
	subordinate clauses	35.16	32.36	26.14
subordinate/main clauses ratio		0.54	0.48	0.36
average depth of 'chains' of embedded subordinate clauses		1.25	1.17	1.01

In spite of the fact that subordination is typically taken as an index of structural complexity, as Mortara Garavelli (2003: 6-8) observed for the Italian legal texts, the use of hypotactic structures can be justified when they allow making plain the hierarchical order of pieces of discourse information otherwise hardly comprehensible. On the contrary, *horizontal* coordinated constructions may cause a conceptual burden not smaller than that of a *vertical* hypotactic structures. While coordinate connectives flatten the hierarchical organization of the discourse, an embedded subordinate construction allows keeping the ordered distribution of pieces of information, e.g. the cause/effect order, exceptions, conditions, etc. Therefore, the higher use of subordination in the legal texts might suggest that the logical structuring of legal discourse is typically expressed through hypotactic structures organised in a hierarchy.

Features concerning nominal modification

The interest in investigating these features stems from Mortara Garavelli's statement that legal sentences are characterised by embedding constructions of nominal modifiers that are typically prepositional complements (Mortara Garavelli, 2001: 171-175). According to her view, such syntactic behaviour causes *structural complexity* of sentences which can affect the *transparency and understandability* of legal documents.

Consider the following sample sentence extracted from the legal corpus:

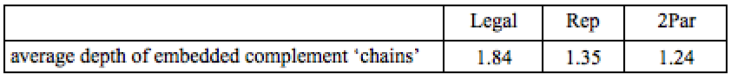
(5) Il Consiglio è giunto ad un **accordo** *sui contributi dei singoli Stati membri all'adempimento dell'impegno globale di riduzione delle emissioni della Comunità* nelle conclusioni del Consiglio del 16 giugno 1998. ('The Council **agreed** *upon the contributions of each Member State to the overall Community reduction commitment* in the Council conclusions of 16 June 1998.')

In (5), the noun *accordo* (the verb 'agreed' in the English translation) is modified by a sequence of 6 embedded prepositional dependency links[7].

The results reported in Table 10 provide empirical validations of Mortara Garavelli's theoretical claims: the legal corpus is characterised by significantly longer complement 'chains'.

The legal corpus is also characterised by a higher percentual distribution of deeply embedded sequences of prepositional complements. In particular, legal texts are char-

Table 10. Nominal modification in legal and newswire corpora.



	Legal	Rep	2Par
average depth of embedded complement 'chains'	1.84	1.35	1.24

acterised by a lower percentage of sequences including one prepositional complement (53.41%) with respect to "La Repubblica" (73.32%) and "Due Parole" (79.40%), and by longer sequences including up to 6 complements. For example, 'chains' of 3 complements constitute 11.80% of the total amount of prepositional complement 'chains' occurring in the legal corpus while they have a coverage of only 4.64% in "La Repubblica" corpus and 2.71% in "Due Parole"; in addition the legal texts are characterised by 5.23% of sequences including 4 prepositional complements while in "La Repubblica" they constitute 0.99% and 0.48% in "Due Parole".

Different types of legal texts in comparison

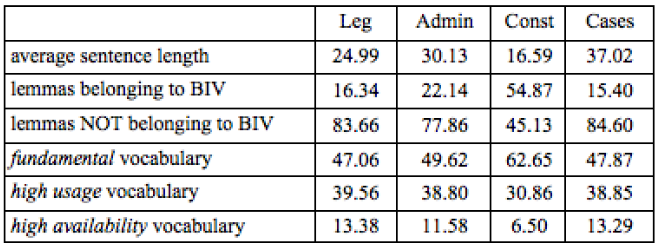
The in depth analysis of the linguistic peculiarities of the different types of legal documents is illustrated in what follows. It aims at highlighting, on the one hand, the linguistic variation between documents used in a "legislative" and "juridical setting", and, on the other hand, the documents enacted or resolved by different authorities.

Different sub-genre of legal texts

The linguistic profiling of the legal documents used in a "legislative setting" (i.e. legislative and administrative acts as well as the Italian Constitution) and in a "juridical setting" (i.e. legal cases) has shown that they differ significantly in many respects at the level of raw and lexical features.

As Table 11 shows, the legal cases (*Cases*) with the longest sentences and the lowest percentage of lemmas (types) belonging to the "Basic Italian Vocabulary" (BIV) result to be the legal sub-genre most different from newspapers. On the contrary, the Italian Constitution (*Const*) is the most similar to them. It contains the highest percentage of words belonging to BIV as well as of the *fundamental* vocabulary and, interestingly, the shortest sentences – even shorter that "Due Parole" corpus (19.20). This is due to the fact that the Constitution witnesses the linguistic efforts of the founding fathers towards a simple and plain legislative drafting in principle comprehensible to a wide public of readers (De Mauro, 2006).

Table 11. Raw and lexical features in sub-genres of legal texts.

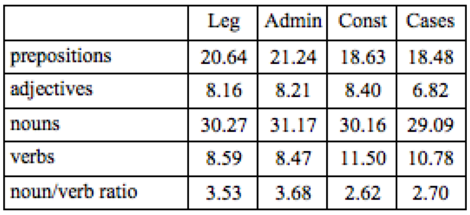


	Leg	Admin	Const	Cases
average sentence length	24.99	30.13	16.59	37.02
lemmas belonging to BIV	16.34	22.14	54.87	15.40
lemmas NOT belonging to BIV	83.66	77.86	45.13	84.60
fundamental vocabulary	47.06	49.62	62.65	47.87
high usage vocabulary	39.56	38.80	30.86	38.85
high availability vocabulary	13.38	11.58	6.50	13.29

If we consider the distribution of the morpho-syntactic characteristics, a number of variations can be observed. In particular, we can see that the administrative (*Admin*) and legislative (*Leg*) documents have a higher percentage of nouns and a lower percentage of verbs than the legal cases. This affects the different values of the noun/verb ratio which is higher in *Admin* and *Leg* than in *Cases*. Following Biber's (1993) outcomes, these results suggest that we are dealing with different language varieties possibly pursuing different communicative purposes. We might put forward here the hypothesis that this provides empirical evidence of the acknowledged difference between the communicative purpose of the documents used in a "legislative setting" and in a "juridical setting". The first ones are in their nature *performative* but they also have distinguishable characteristics of *informative* texts since "every attempt is made to write not only clearly, precisely and unambiguously but also all-inclusively" (Bathia, 1987: 230). Conversely, legal cases serve several different communicative purposes (Bathia, 1993). In particular, the threefold internal structure of Italian legal cases, which is overtly established by article 118 of the Italian Civil Procedure Code, corresponds to three different communicative purposes (Santulli, 2008): *narrative* (corresponding to the legal case section where the facts which are relevant for the case are reported), *argumentative* (the function of the section where the judge reports the motivations of the final decision) and *performative* (the function of the last section, i.e. the final decision).

This difference may affect the different noun/verb ratio values reported here.

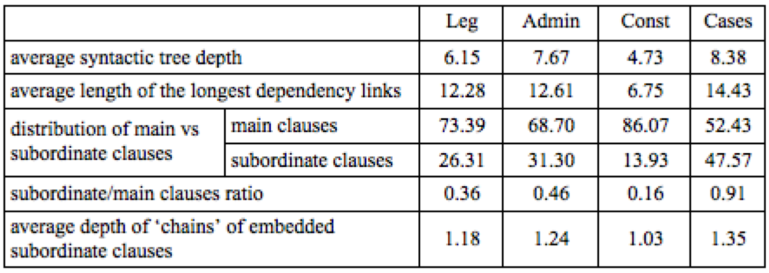
Table 12. Morpho-syntactic features in sub-genres of legal texts.



	Leg	Admin	Const	Cases
prepositions	20.64	21.24	18.63	18.48
adjectives	8.16	8.21	8.40	6.82
nouns	30.27	31.17	30.16	29.09
verbs	8.59	8.47	11.50	10.78
noun/verb ratio	3.53	3.68	2.62	2.70

Moving to the analysis of the syntactic features, it results that the corpus of legal cases contains the deepest syntactic trees and the longest dependency links (see Table 13), i.e. two of the features mostly indicative of syntactic complexity (see Section 'Features based on the structure of a syntactic tree'). Conversely, the 'easiest' structures occur in the Constitution which shows values lower also with respect to "Due Parole" overtly written in a plain language. Among the documents used in a "legislative setting", the administrative acts result to be the most complex ones.

Table 13. Syntactic features in sub-genres of legal texts.



		Leg	Admin	Const	Cases
average syntactic tree depth		6.15	7.67	4.73	8.38
average length of the longest dependency links		12.28	12.61	6.75	14.43
distribution of main vs subordinate clauses	main clauses	73.39	68.70	86.07	52.43
	subordinate clauses	26.31	31.30	13.93	47.57
subordinate/main clauses ratio		0.36	0.46	0.16	0.91
average depth of 'chains' of embedded subordinate clauses		1.18	1.24	1.03	1.35

The four legal sub-corpora differ greatly with respect to the percentage distribution of the subordinate clauses. In particular, the legal cases contain the highest percentage of subordinate clauses organised in deep ‘chains’: sequences of e.g. 2 embedded subordinate clauses constitute 20.58% of the total amount of ‘chains’ of subordinate clauses in this legal sub-genre while they are 15.75% in *Admin*, 12.54% in *Leg* and only 3.17% in *Const*. Consequently, legal cases have the highest subordinate/main clauses ratio. Interestingly, if we focus on the documents used in a “legislative setting”, we can see that the Constitution shows values lower also with respect to “Due Parole” while the administrative acts contain the highest percentage of subordinate clauses organised in longer sequences with respect to the legislative texts.

The *Leg* corpus stands out for the greater use of nominal modification. In particular, it contains the highest percentual distribution of deep embedded sequences of prepositional complements. ‘Chains’ of e.g. 3 complements constitute 12.07% of the total amount of prepositional complement ‘chains’ while they are 11.10% in *Cases*, 9.77% in *Admin* and 5.67% in *Const*.

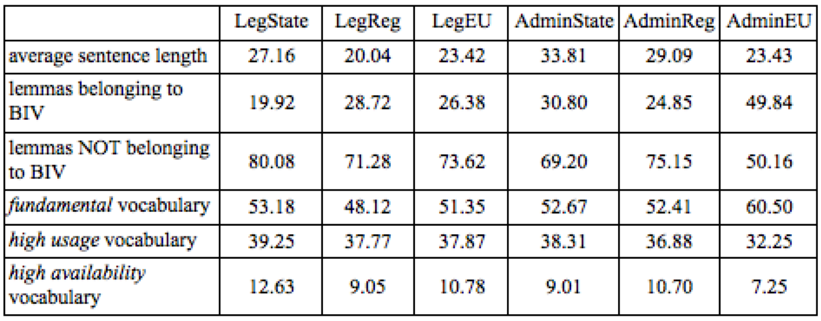
Legal texts enacted or resolved by different authorities

The “legislative setting”

Significant linguistic variation can be observed by comparing legal documents used in the same setting but released by different authorities. Starting from the analysis of the documents used in the “legislative setting”, acts enacted by the European Commission, the Italian State or by the Piedmont Region differ significantly at the level of raw and lexical features.

The national (*AdminState* in all tables) and regional (*AdminReg*) administrative acts not only have the longest sentences but they also contain the lowest percentage of lemmas (types) belonging to the “Basic Italian Vocabulary” (BIV) and of fundamental vocabulary (see Table 14). The European administrative texts (*AdminEU*) represent the opposite pole with the shortest sentences and the highest percentage of BIV. A similar variation between national (*LegState*) and European (*LegEU*) legislative texts can be observed, even if regional legislative acts (*LegReg*) do not follow this trend.

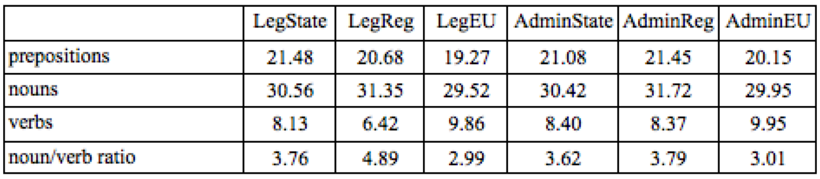
Table 14. Raw and lexical features in legislative and administrative texts.



	LegState	LegReg	LegEU	AdminState	AdminReg	AdminEU
average sentence length	27.16	20.04	23.42	33.81	29.09	23.43
lemmas belonging to BIV	19.92	28.72	26.38	30.80	24.85	49.84
lemmas NOT belonging to BIV	80.08	71.28	73.62	69.20	75.15	50.16
fundamental vocabulary	53.18	48.12	51.35	52.67	52.41	60.50
high usage vocabulary	39.25	37.77	37.87	38.31	36.88	32.25
high availability vocabulary	12.63	9.05	10.78	9.01	10.70	7.25

Moving to the analysis of the morpho-syntactic features, both administrative and legislative acts enacted by the Italian State and the Piedmont Region result to be characterised by a higher percentage of prepositions and nouns with respect to the European documents and by a lower percentage of verbs (see Table 15). This affects the different noun/verb ratio (see Table 15).

Table 15. Morpho-syntactic features in legislative and administrative texts.

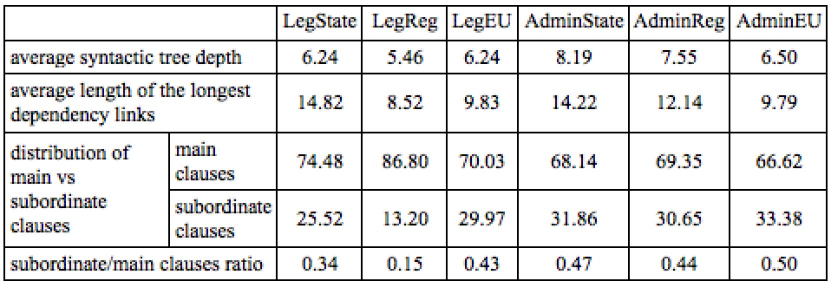


	LegState	LegReg	LegEU	AdminState	AdminReg	AdminEU
prepositions	21.48	20.68	19.27	21.08	21.45	20.15
nouns	30.56	31.35	29.52	30.42	31.72	29.95
verbs	8.13	6.42	9.86	8.40	8.37	9.95
noun/verb ratio	3.76	4.89	2.99	3.62	3.79	3.01

National and regional administrative acts resulted to be more syntactically complex than the European acts: they have deeper syntactic trees and longer dependency links (see Table 16). Similar to the national and regional legislative acts, they contain longer sequences of embedded prepositional complements: in *LegEU* and *AdminEU* 'chains' of e.g. 3 complements constitute 11.18% and 7.98% respectively of the total amount while they a coverage of 12.54% in *LegState*, 12.24% in *LegReg*, 9.71% in *AdminState* and 9.99% in *AdminReg*.

A distinguishing characteristic of the European documents is the higher occurrence of subordinated constructions with respect to national and regional documents. As Table 16 shows, both legislative and administrative European legal documents are characterised by a higher percentage of subordinate clauses and by a higher subordinate/main clauses ratio. As discussed in Section 'Features concerning the use of subordination', these results should be combined with the study of how pieces of discourse information are logically organized throughout a document. However, this syntactic behaviour confirms a tendency already observed in this study: the linguistic similarity of the European legal language to ordinary language is greater than the language variety used in the documents enacted by the Italian State and the Piedmont Region.

Table 16. Syntactic features in legislative and administrative texts.



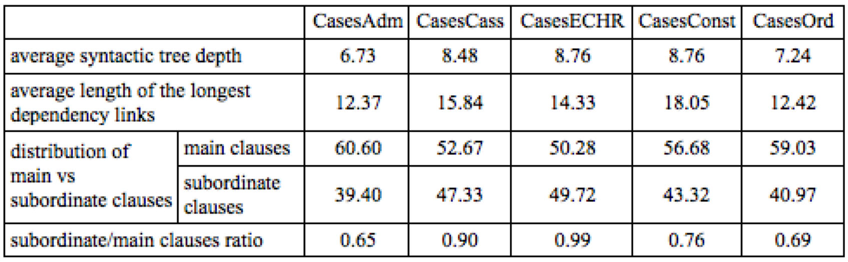
		LegState	LegReg	LegEU	AdminState	AdminReg	AdminEU
average syntactic tree depth		6.24	5.46	6.24	8.19	7.55	6.50
average length of the longest dependency links		14.82	8.52	9.83	14.22	12.14	9.79
distribution of main vs subordinate clauses	main clauses	74.48	86.80	70.03	68.14	69.35	66.62
	subordinate clauses	25.52	13.20	29.97	31.86	30.65	33.38
subordinate/main clauses ratio		0.34	0.15	0.43	0.47	0.44	0.50

The "juridical setting"

A number of different linguistic peculiarities can be noted in the five sub-corpora of legal cases. As Table 17 shows, the documents issued by the Constitutional Court (*CasesConst* in all tables) contain the deepest syntactic trees and the longest dependency links while the legal cases resolved by Ordinary Tribunals (*CasesOrd*) and Administrative Courts (*CasesAdm*) have an opposite behaviour. Moreover, the *CasesConst* corpus contains the highest percentual distribution of deeply embedded sequences of prepositional complements. 'Chains' of 3 complements e.g. constitute 9.84% of the total amount while they have a coverage of 9.27% in *CasesCass*, 8.14% in *CasesECHR*, 7.98% in *CasesOrd* and of 7.56% in *CasesAdm*.

Moving to the analysis of the use of subordination, it resulted that the documents issued by the Court of Civil Cassation (*CasesCass*) and by the European Convention on Human Rights Court (*CasesECHR*) are characterised by the highest occurrence of subordinated constructions. Even if, according to the literature on the topic, these results can be interpreted in various manner (see Section 'Features concerning the use of subordination'), they confirm the linguistic differences between Ordinary Tribunals and Administrative Courts cases, on the one side, and the other types of legal cases, on the other side.

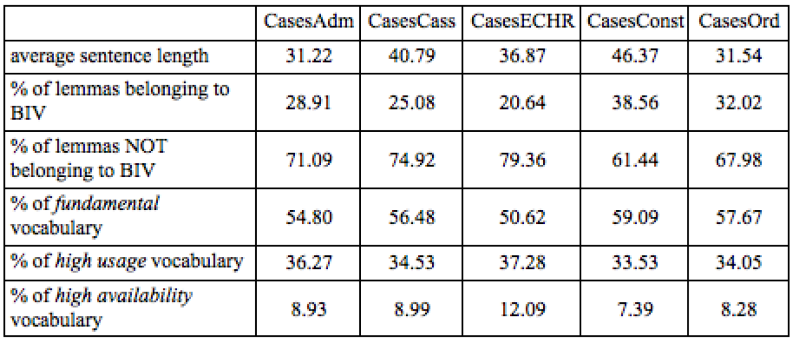
Table 17. Syntactic features in legal cases.



		CasesAdm	CasesCass	CasesECHR	CasesConst	CasesOrd
average syntactic tree depth		6.73	8.48	8.76	8.76	7.24
average length of the longest dependency links		12.37	15.84	14.33	18.05	12.42
distribution of main vs subordinate clauses	main clauses	60.60	52.67	50.28	56.68	59.03
	subordinate clauses	39.40	47.33	49.72	43.32	40.97
subordinate/main clauses ratio		0.65	0.90	0.99	0.76	0.69

Interestingly, the corpus of Constitutional Court cases is characterised by the highest percentage of lemmas belonging to the "Basic Italian Vocabulary" (BIV) and by the highest percentage of *fundamental* vocabulary (see 18). Apparently inconsistent with the literature on text complexity, this demonstrates, on the contrary, how an exhaustive linguistic investigation should consider the complex interaction of different kinds of linguistic features. This result tells us that in *CasesConst* a basic vocabulary is used in complex syntactic constructions occurring in long sentences while in *CasesAdm* a more 'complex' vocabulary is used in less complex syntactic structures occurring in short sentences.

Table 18. Raw and lexical features in legal cases.



	CasesAdm	CasesCass	CasesECHR	CasesConst	CasesOrd
average sentence length	31.22	40.79	36.87	46.37	31.54
% of lemmas belonging to BIV	28.91	25.08	20.64	38.56	32.02
% of lemmas NOT belonging to BIV	71.09	74.92	79.36	61.44	67.98
% of *fundamental* vocabulary	54.80	56.48	50.62	59.09	57.67
% of *high usage* vocabulary	36.27	34.53	37.28	33.53	34.05
% of *high availability* vocabulary	8.93	8.99	12.09	7.39	8.28

Conclusion

In this paper, the author presented an NLP-based study aimed at performing the linguistic profiling of a corpus of different types of Italian legal texts exemplifying different sub-varieties of Italian legal language. She analysed the distribution of a wide range of different linguistic features automatically extracted from text in order to investigate the linguistic variation between *i)* the considered corpus of legal texts and a corpus of newspaper articles representative of Italian ordinary language and *ii)* between different types of legal texts that have been compared at different levels of specificity.

The followed comparative approach has allowed the investigation of lexical, morpho-syntactic and syntactic characteristics which make the corpus of legal texts different from newspaper articles. Interestingly, the legal language resulted closer to the ordinary language used in "La Repubblica" corpus than in "Due Parole" corpus. It is particularly the case when we took into consideration the low percentage of lemmas belonging to the "Basic Italian Vocabulary" or features typically taken as indices of syntactic complexity such as syntactic tree depth and length of dependency links. This shows that it is very often the case that legal texts do not conform to the suggestions put by the "Guide to drafting administrative acts", the most up-to-date guide describing the lexical, morpho-syntactic and syntactic characteristics that a legal document is expected to have in order to be written in a *plain, simple and comprehensible language.*

The present study has also highlighted systematic differences between the considered sub-varieties of the legal language. The 'genre-internal' perspective of analysis adopted here has shown that significant linguistic variations exist not only among documents used in different settings but also among documents enacted or resolved by different authorities. It has been demonstrated that the Italian Constitution articles were written using a legal language variety very close to the language of "Due Parole" corpus that was specifically written using a plain and controlled language. This empirically witnesses the linguistic efforts of the founding fathers towards a simple and plain legislative drafting. We have also discussed which are the linguistic characteristics making the legal texts enacted by the European Commission more similar to the ordinary Italian than the acts enacted by the Italian State and the Piedmont Region. Finally, it has been shown that among the legal cases the ones resolved by the Constitutional Court have the greater number of features typically taken as indices of syntactic complexity.

If on the one hand the approach to the linguistic profiling proposed here has been devoted to showing how computational linguistic analysis techniques can help to shed light on some main peculiarities of the Italian legal language, providing quantitative validations of theoretical claims from the literature, on the other hand, different types of applications could benefit from the results of this study. They can be used as a starting point to identify areas of lexical, morpho-syntactic and/or syntactic complexity within a legal text and/or a single sentence in order to assess their readability. Similarly, the investigation of the wide range of linguistic characteristics carried out in this study might be exploited to perform a number of different computational forensic linguistics tasks – first and foremost authorship attribution.

Acknowledgments

The research reported in the paper was supported by a research grant through the Scuola Superiore Sant'Anna in Pisa. The author would like to thank in particular professor Antonio Lazari who directed the construction of the corpus of legal cases analysed in this study.

The linguistic profiling methodology followed here represents an essential line of research of the ItaliaNLP Lab[8] at the Institute of Computational Linguistics "Antonio

Zampolli" (ILC-CNR). The author greatly acknowledges the help and contributions in particular of Simonetta Montemagni and Felice Dell'Orletta whose invaluable comments and ideas helped to carry out the work presented here.

Notes

[1] http://www.dueparole.it/

[2] The annotation format adheres to the standard CoNLL-2007 tabular format used in the "Shared Task on Dependency Parsing" (Nivre *et al.*, 2007).

[3] Evalita, 2009

[4] The Italian version of the "Guide" is available at http://www.pacto.it/content/view/416/48/

[5] In Figure 1, each consecutive dependency link is highlighted with a frame.

[6] Note that also punctuation is included.

[7] The prepositions heading the prepositional complement are underlied here. Note that the complement *dei singoli Stati Membri* ('of each State Member') has not been included in the embedded sequence since it modifies the noun *contributi* ('contributions').

[8] http://www.italianlp.it/

References

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, 166–170, New York City, New York.

Bathia, V. K. (1987). Language of the law. *Language Teaching*, 20(4), 227–234.

Bathia, V. K. (1993). *Analysing genre. Language Use in Professional Settings.* London and New York: Longman.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen and R. Freedle, Eds., *Coherence in Spoken and Written Discorse*, 45–80.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics Journal*, 19(2), 219–241.

Biber, D. and Conrad, S. (2009). *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Cortelazzo, M. (1997). Lingua e diritto in Italia. il punto di vista dei linguisti. In L. Schena, Ed., *La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno Internazionale*, Milano: Roma, Cisu (Centro d'Informazione e Stampa Universitaria).

De Mauro, T. (2006). Introduzione. Il linguaggio della Costituzione. In *Costituzione della Repubblica Italiana (1947).* Torino: UTET.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia.

Dell'Orletta, F., Marchi, S., Montemagni, S., Plank, B. and Venturi, G. (2012a). The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts. In *Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts"*, 42–51.

Dell'Orletta, F., Marchi, S., Montemagni, S., Venturi, G., Agnoloni, T. and Francesconi, E. (2012b). Domain Adaptation for Dependency Parsing at Evalita 2011. In *Working Notes of EVALITA 2011.*

Dell'Orletta, F. and Montemagni, S. (2012). Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico. In S. Ferreri, Ed., *Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI)*, 343–359, Roma.

Dell'Orletta, F., Montemagni, S., Vecchi, E. M. and Venturi, G. (2011a). Tecnologie linguistico-computazionali per il monitoraggio delle competenze linguistiche di apprendenti l'italiano come L2. In G. C. Bruno, I. Caruso, M. Sanna and I. Vellecco, Eds., *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, 319–336. Milano: McGraw–Hill.

Dell'Orletta, F., Montemagni, S. and Venturi, G. (2011b). READ-IT: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, 73–83.

Dell'Orletta, F., Montemagni, S. and Venturi, G. (2012c). Genre-oriented Readability Assessment: a Case Study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education (SLP-TED)*, 91–98.

Dell'Orletta, F., Montemagni, S. and Venturi, G. (2013). Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, 189–197, Hissar.

Fornaciari, T. and Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen and A. M. Zwicky, Eds., *Natural Language Parsing*, 129–189. Cambridge: Cambridge University Press.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76.

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing*, 167–202.

Givón, T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, 15(2), 335–370.

Hudson, R. (1995). Measuring syntactic difficulty. Unpublished paper.

Lazari, A. (2005). *Modelli e Paradigmi della Responsabilità dello Stato.* Torino: Giappichelli.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on pur capacity for processing information. *Psycological Review*, 63, 81–97.

Montemagni, S. (2013). Tecnologie linguistico-computazionali e il monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata.*

Mortara Garavelli, B. (2001). *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani.* Torino: Einaudi.

Mortara Garavelli, B. (2003). Strutture testuali e stereotipi nel linguaggio forense. In P. M. Biagini, Ed., *La lingua, la legge, la professione forense. Atti del convegno Accademia della Crusca*, 3–19,, Milano: Giuffrè.

Nivre, J., Hall, J., Kubler, S., McDonald, R., Nilsson, S., Riedel, S. and Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the EMNLP–CoNLL*, 915–932.

Piemontese, M. E. (1996). *Capire e farsi capire. Teorie e tecniche della scrittura controllata.* Napoli: Tecnodid.

Rovere, G. (2005). *Capitoli di linguistica giuridica. Ricerche su corpora elettronici.* Alessandria: Edizioni dell'Orso.

Santulli, F. (2008). La sentenza come genere testuale: narrazione, argomentazione, performatività. In G. Garzone and F. Santulli, Eds., *Linguaggio giuridico e mondo contemporaneo*, 207–238. Milano: Giuffrè.

Sousa-Silva, R., Sarmento, L., Grant, T., Oliveira, E. and Maia, B. (2010). Comparing Sentence-Level Features for Authorship Attribution in Portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, 51–54.

van Halteren, H. (2004). Linguistic Profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics*, 200–207.

Venturi, G. (2012). Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts. In *Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts"*, 1–12.