IN PRESS FOR Proceedings of the workshop "*Forensic linguistics between scientific research and legal practice", LIV Conference of the Italian Society for Linguistics (SLI)*

LUCIA BUSSO
ASTON INSTITUTE FOR FORENSIC LINGUISTICS, ASTON UNIVERSITY
BIRMINGHAM, UK

***CorIELLS*: a specialised bilingual corpus of English and Italian legal-lay language**

Abstract: This contribution presents *CorIELLS* (Corpus of Italian and English Legal-Lay textS). *CorIELLS* is an open-access bilingual corpus of legal-lay language for Italian and English. The corpus rationale, collection, and composition are discussed, together with previous research on the corpus in both Italian and English. Furthermore, a cross-linguistic exploration of lexical and syntactical linguistic complexity is presented, using the *Profiling-UD* tool (Brunato et al., 2020) is further presented. Dimensions of complexity are investigated with Principal Component Analysis (Kassambara, Mundt 2020). Results for the two languages are compared, and similarities and differences in dimensions of complexity are foregrounded.

Keywords: corpus linguistics; legal-lay language; language complexity

1. *Introduction*

The present contribution describes a new linguistic resource for the analysis of an under-researched genre: legal-lay language. We define here this textual type as the language of any type of document with legal content aimed at non-specialist audiences (Crystal, Davy, 1969; Tiersma, 1999). Legal-lay language (henceforth: LLL) is increasingly important in our everyday life. Particularly, one constantly finds oneself in need of understanding some sort of legal document, from consumer contracts to privacy notes and terms and conditions of websites.

Since research on LLL as an independent genre is still scarce, this contribution presents the first specialised corpus on LLL, *CorIELLS*, and illustrates some types of research that can be conducted on it. Specifically, the remaining of section 1 (sections 1.1. and 1.2) outline more in depth the background literature and the motivations of the research. In section 2 we describe the corpus collection, its rationale, and the corpus composition. Section 3 outlines previous research that was conducted using the CorIELLS (3.1) and presents tan innovative analysis of linguistic complexity using dimensionality reduction (3.2).

The discussion in section 4 concludes the paper, summarizing the core notions presented and opening new lines of possible research.

1.1. Why legal-lay language?

It is widely acknowledged in the scientific literature that lack of understanding of legal-lay language leads to problems legislative and linguistic alike (Gibbons, 2003; Curtotti, McCreath, 2013; Benoliel, Becher, 2019). The many issues arising from an incomplete understanding of LLL has been foregrounded by many scholars. For example, Tiersma (1999:20) claims that "people have the right to know the meaning of the contracts that they sign and for which they will be held legally responsible. When people are entitled to understand a legal document, it should be as free as possible of technical terms and jargons. If technical terms are avoidable, they should at least be explained in ordinary language". Frade (2007, 2016) also identifies asymmetrical power dynamics between powerful business companies and their customers, with the formers holding a "hidden power" (p.1). This notion is cognate to Fairclough's (1995) concept of 'technologization of discourse' – i.e., the manipulation of social practices by the more powerful social force.

Related to this issue, consumers are not keen or motivated in reading the contract provisions or legal notice documents, thus committing themselves to an often legally binding relation they are not aware of.

Despite the centrality of LLL in the contemporary world, studies dedicated specifically to this textual type specifically are not common in the literature (Brunato, Venturi 2014; Lintao, Madrunio, 2015; Van Boom et al., 2016; Conklin et al., 2019). In fact, the scientific literature often assumes that LLL is not a textual type independent from legal language, but only its 'simplified' and more accessible version (Bhatia, 1983; Venturi, 2011; Adler, 2012). Particularly, the *Plain Language movement* (Williams, 2004; Adler, 2012) and the Italian version *Progetto Chiaro* (Williams, 2005) have been advocating for decades the need for a plainer language in drafting legal documents directed at ordinary citizens. This is a difficult balance to obtain, as LLL must crucially preserve its intended legal meaning (Kimble, 2000; Eagleson, 2004) while at the same time be drafted in a language plain enough for non-legal specialists to read and understand.

We argue that this delicate balance between legal content and plainer form constitute *de facto* a novel textual type, autonomous and independent from its 'parent' genre, legal jargon. Although the precise connotations of the concept 'genre' are eluding and there is no consensus in the literature, one of the most famous definitions of it states that:

> "genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognised by the expert members of the parent discourse community, and thereby

> constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style." (Swales 1990: 58)

Similarly, Bhatia (2004: 23) notes that

> "[g]enre essentially refers to language use in a conventionalised communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms by imposing constraints on the use of lexicogrammatical as well as discoursal resources."

Therefore, we argue that the differences in audience and community of practice, the different aims, the conscious effort to simplify both lexicon and grammar, and the different medium (given that LLL is mainly found on the Internet or in consumer contracts) makes it reasonable to consider LLL as an idiosyncratic genre (Busso, 2022; Busso, forthcoming).

Given the scarcity of linguistic studies on LLL as a separate entity from legal language in the literature and the absence of dedicated linguistic resources, the present paper presents the first specialised bilingual corpus of LLL: *CorIELLS* (Corpus of Italian and English Legal-Lay textS). The corpus was originally collected to investigate linguistic complexity in LLL, but it also lends itself to different analyses. Corpus files are freely available from download from REDACTED FOR REVIEW.

1.2. Why a legal-lay language corpus?

As argued by several scholars in the literature, compiling relatively small and specialised corpora – instead of mega corpora of billions of words – allows researchers to investigate fine-grained aspects of language, like a specific register or genre (Flowerdew 2002; Hyland, 2006; Koester, 2010). In fact, while many larger corpora are suitable to research general linguistic phenomena and aim to represent a larger variety of language (e.g., British English), specialised corpora are generally catered to more specific research questions. In specialised corpora,

> "much closer link between the corpus and the contexts in which the texts in the corpus were produced. Where very large corpora, through their de-contextualisation, give insights into lexico-grammatical patterns in the language as a whole, smaller specialised corpora give insights into patterns of language use in particular settings" (Koester, 2010:67)

Therefore, to investigate features of LLL a new corpus was collected, following Flowerdew (2004) guidelines to build specialised corpora for specific genres and Biber (1993)'s insights on situational (i.e., the range of text types) and linguistic (i.e., the range of linguistic distributions) representativeness. By including a wide variety of text types in the corpus (see section 2 below), the aim is to yield insights not only into the particular corpus, but also into more vast language use in the genre of LLL (Tognini Bonelli, 2001).

2. *The CorIELLS corpus and its use*

The corpus was compiled as to pe partially parallel between Italian and English. That is, documents included in the final corpus selection are both comparable (i.e., same content in the two languages) and idiosyncratic texts (i.e., same document type but different content). The part of the corpus designed to be parallel only includes document drafted by two different legal teams in the two languages. This prevents – to the best of the author's possibilities – spurious phenomena related to specialised translation (Matulewska, 2007; Biel, 2009, 2018). Moreover, many scholars have claimed that different versions of the same text can be considered autonomous texts, as they are equally "authentic but not necessarily based on translations from one or the other" (Williams, 2004b: 219). The independence of translated texts is also advocated from within translation studies, as translations should be regarded as "autonomous texts destined to function in the context of the target culture without regard for their relation to the source text" (Garzone, 2008:48).

The final selection of text types includes four major categories of documents: standard legal notices for bank accounts, summaries of European legislation, terms and conditions of websites, standard consumer contracts for utilities (mobile phone contracts, gas, and electricity). Table 1 outlines corpus composition in the two subcorpora of Italian and English texts.

The corpus was collected semi-automatically with the web-scraping toolkit *Bootcat* (Baroni, Bernardini, 2004).

Final corpus size amounts to over 1.6M words. Despite being relatively small and highly specialized in nature, *CorIELLS* aims at obtaining a representative sample of the genre in question (Biber, 1993; Koester, 2010), while maintaining as a strict parameter of inclusion only freely accessible online texts. In this way, we approximate the types of LLL than any ordinary person could obtain by browsing the web.

Section 2.1 will describe in more detail the selection and collection procedure for each subsection, while section 2.2 will present previous work using the corpus as background to the study here described in section 3.

Document type	Document nº	English subcorpus (740K words)	Italian subcorpus (880K words)	Comparable
Bank account contracts	15 in total	19%	27%	no
Utilities contracts	15 per language	25%	23%	no
EurLex summaries	247 per language	22%	23%	yes
Terms and conditions	45 per language	34%	27%	yes

Table 1 - *composition of the CorIELLS corpus*

2.1. Corpus selection

As mentioned, four major categories of documents compose CorIELLS (table 1). The composition of the corpus is by no means exhaustive of the different types of texts that can be included under the umbrella term of LLL. The final selection represents a representative sample of document types freely available online. All document types that it was not possible to retrieve freely were automatically excluded. This criterion is underpinned by the assumption that if a document is not available for download/ view on the open web, it may very well only shown by professionals explaining the wordings and the main clauses and provisions (for example, insurance contracts). Table 2 below reports all document sources for the corpus.

The selected documents are:

1. Standard legal notices for bank current accounts. A selection of 15 common bank institutes in both Italy and the UK were chosen (7 for Italian and 8 for English, given the different lengths of the documents). As mentioned, only standard contracts for current accounts freely available online were selected. The documents were first retrieved in their PDF form and text files were obtained using *Bootcat*.
2. Utilities standard contracts. For this section, 15 widely common companies for every-day utilities in Italy and the UK were selected. Standard contract terms for 5 energy supply (gas and electricity), 5 Wi-Fi suppliers, and 5 pay-by-month phones were downloaded as PDF and transformed into text files using *Bootcat*.
3. Terms and conditions (or terms of use) of websites. To have a representative and balanced set of data, the list of the 500 most visited websites in Italy and the UK in 2019 was used. We then selected websites that were present in both lists and only web services with legal notices in both languages were kept. Furthermore, the websites were manually checked to make sure that the two legal notices had been drafted by two legal teams. When terms of use and terms and conditions were in different pages, both were selected. A final selection of 45 websites were included in the corpus. Text files were extracted

from the URLs of the terms and conditions pages using the URL collection function on *Bootcat.*

4. European legislation summaries. These texts were retrieved from the EUR-Lex website (https://eur-lex.europa.eu/homepage.html). Summaries are defined as "short, easy-to-understand explanations of the main legal acts passed by the EU–intended for a general, non-specialist audience" *(EUR-lex* website). All summaries from 2019 and 2020 (as of December 1st, 2020) were retrieved from the official website (using the Advanced Search function), in both their Italian and English version. These documents are originally drafted in English and later adapted by specialised translators and legal experts in each language of the European Union, as prescribed in EU style guides (Interinstitutional Style Guide, 2015:54-62). In this case as well, URLs were inputted into *Bootcat* to extract their plain text version.

Corpus section	**Documents**
Bank contracts	**ITA:** Illimity, BPM, Finecobank, Banca Mediolanum, Deutsche Bank, Banca Sella, Banca d'Italia. **ENG:** RBS, N26, Starling Bank, Monzo, HSBC, Lloyds, Nationwide, Revolut.

Utilities contracts	**ITA:**
	Energy: Enel Energia, Alleanza Luce&Gas, AGSM, Illumia, Iren;
	Phone: Vodafone IT, Tim, Iliad, Wind, Tre;
	Wi-fi: Tiscali, Tim ADSL, Vodafone Casa, Wind ADSL, Fastweb.
	ENG:
	Energy: Southern Electric, EDF Energy, Gazprom, Go Power, Npower;
	Phone: EE, GiffGaff, O2, Tesco Mobile, Vodafone UK;
	Wi-fi: NOW TV, BT broadband, SKY UK, Vodafone Broadband

T&C	AirBNB (terms of use + t&c), Alibaba (terms of use + t&c), Amazon, Asos, Booking, DropBox, EasyJet, Ebay, Etsy, Europeancommission, Facebook, Fandom, Firefox, Google, Groupon, H&M, HP, Ikea, Instagram, JustEat, Linkedin, LiveJasmin, Livesport, Microsoft, Netflix, Pinterest, Primevideo, Ryanair, Samsung (terms of use + t&c), Shein, Skyscanner, Spotify, Tripadvisor, Trustpilot, Tumblr, Twitch, Twitter, Vice, WhatsApp, Wikimedia Foundation, Wikipedia, Yahoo, Youtube.
EurLEX summaries	All summaries from January 2019 to November 2020 at https://eur-lex.europa.eu/eu-summary/eu-summary-search.html

Table 2 – *Documents retrieved in corpus selection*

As mentioned, the corpus files are freely available from download from the Forensic Linguistic Databank (FoLD, https://fold.aston.ac.uk/), an innovative online repository hosted by Aston University[1].

2.2. Previous research on legal-lay language

A couple of studies (Busso, 2022, forthcoming) have used CorIELLS to investigate different aspects of LLL in terms of its lexico-grammatical features and its comprehensibility.

Specifically, Busso (2022) presents a quantitative text-based analysis rooted in the core principles of Construction Grammar (Goldberg, 2006, 2019). The study analyses four constructions typical of Italian and English legal jargon and LLL (Mortara-Garavelli, 2001; Williams, 2004; Chovanec, 2013; Mori, 2019). The same constructions are also used for the study in Busso (forthcoming):

1. Modal verbs (MOD)
    a. *I paesi UE devono notificare alla Commissione eventuali obblighi e requisiti in materia di comunicazione* (EU countries must notify the Commission of any communication obligations and requirements)
    b. We must be satisfied of your identity and can refuse instructions if we doubt your identity
2. Nominalisations heading PP attachment chains NOM-pp
    a. *Informazioni sul sistema di risoluzione delle controversie di cui alla Delibera Consob* (Information on the system of resolution of litigations referred to in the Consob resolution)
    b. Mandatory collective management of rights for retransmissions of radio and television programmes by means other than cable.
3. Reduced participial relatives PART
    a. *Nota informativa concernente il trattamento e la protezione dei dati personali.* (Information note treating the processing and protection of personal data)
    b. The 'application publisher' means the entity licensing the application to you as identified in the Store.
4. Passive constructions PASS
    a. *Il reclamo può essere presentato anche dopo la data di entrata in vigore della variazione.* (The complaint can also be submitted after the effective date of the change)
    b. Payments (…) will be sent on the next working day.

[1] FoLD is a "permanent, controlled access online repository for forensic linguistic data" (Petyko et al., forthcoming).

First, a collostructional analysis (Stefanowitsch, 2013) is performed on the selected set of constructions, and statistically significant results are contrasted with the 'Nuovo Vocabolario di Base' (De Mauro, Chiari, 2016). Secondly, results from the first part of the study are further compared to neighbouring genres (i.e., legal jargon and written prose) in a contrastive frequency analysis. These genres are represented with the CORIS corpus (Rossini-Favretti, 2002). Specifically, the legal and narrative subcorpora are used to approximate respectively legal jargon and general domain written prose.

Findings from this study suggest that LLL is a 'blended' genre containing features from both specialistic and non-specialistic registers. The contrastive analysis moreover shows that some lexico-grammatical features are used significantly differently in LLL than in both legal jargon and general-domain written prose. Specifically, all the analysed constructions behave differently in LLL except for NOM-pp (see examples 1-4 below) – which seem to be the most tightly related to the legal genre.

Busso (2022) performs a similar analysis on the English subcorpus of CorIELLS, using the same structures and the New General Service List (Brezina, Gablasova, 2015) as a core vocabulary. Both collostructional analysis and comparative analysis with legal jargon and written prose were performed. For English, CorIELLS is contrasted to an ad-hoc subcorpus of the EurLEX corpus (Baisa et al., 2016) which includes only legislative texts spanning from the 90s to 2015, and to the imaginative subcorpus of the BNC (BNC, 2007).

Findings are comparable with Italian, in that NOM_pp is the most specialist construction out of the four investigated. Moreover, results also show that LLL displays linguistic features quantitatively different from the other two genres. Moreover, a representative sample of concordances is analysed in terms of text-based readability metrics and presented to native speakers in a survey. The comparison between results from these two methods (i.e., readability scores and speakers' judgments) suggests that readability metrics might underestimate the readers' ability to understand LLL texts.

3. *Language complexity in CorIELLS: an exploratory analysis*

As outlined in paragraph 2.2 above, previous research using the CorIELLS corpus has explored the lexico-grammatical level and readability. In the present contribution we expand the existing exploration of this corpus by presenting an analysis of lexico-syntactic complexity of the CorIELLS corpus as compared to legal jargon and written prose, in both Italian and English.

As anyone working with the multifaceted notion of linguistic complexity knows, "complexity" is an ill-defined phenomenon. Many scholars in linguistics have given different definitions and quantitative measures of it (among many others: Benoit, 1990; Kusters, 2003, 2008; Pallotti, 2015). In line with Pallotti (2015), we advocate for a clear-cut working definition of complexity, "treating it as a purely descriptive category, limiting its use to structural complexity and excluding from its definition any theoretical assumption about when, how and why it increases or remains constant" (Pallotti, 2015:119). Therefore, we here denote "complexity" in the most neutral way possible, simply referring to it as a property of language arising from lexico-syntactic and semantic features of it. We also consider the notion of complexity as cognate – although not synonymic – with "comprehensibility".

3.1. Data and methodology

In order to compare the two languages in terms of complexity, a subset of the linguistic parameters provided by the online tool *Profiling-UD* (Brunato et al., 2020) are extracted. This text analysis tool allows for the extraction of a vast number of "linguistic profiling" features across different levels of linguistic annotation. Particularly interesting for the purposes of the present paper is the fact that Profiling-UD is specifically devised to support cross-linguistic analyses, since it is based on the Universal Dependencies (UD) representation (Nivre, 2015).

The comparative analysis was performed using the same corpora of legal and general written language used for previous works: the legal and fiction subcorpus of CORIS for Italian, and EurLEX and the BNC imaginative subcorpus for English (see section 3.1 above). The Italian corpus CORIS was chosen as it is the reference corpus for contemporary written Italian, and reflects a type of Italian variety that can be defined as 'written-written', following Nencioni (1983)'s criteria. It was accessed from the corpus web interface (https://corpora.ficlit.unibo.it/TCORIS/). The fiction subcorpus (CORIS_narr) includes both novels and short stories. The narratives are further subdivided into adult, children, adventure, science-fiction and women literature. Total word count amounts to about 25M words. The legal subcorpus (CORIS_law) is one of the most comprehensive datasets of Italian legal language, including strictly legal, bureaucratic, and administrative texts. It amounts to 10M words.

For English, two separate corpora were used: for legal language, an ad-hoc subcorpus of the *EurLEX* (Baisa et al., 2016) corpus was created, to include legislative documents in English ranging from the 1990s to 2015 (EUR), for comparability reasons; the resulting subcorpus amounts to 600M words. As a proxy for general written language, the "imaginative" subcorpus of the BNC (BNC, 2007) was used, amounting to 20M words. Both corpora were accessed via the *SketchEngine* web interface.

As a first step, a random sample of concordances was extracted from the Italian and English subcorpus of CorIELLS and from the reference subcorpora. The concordances were selected containing the 4 constructions already used for previous analyses (see section 3.1). CQL and CQP searches for the 4 constructions were performed respectively on *SketchEngine* (Kilgarriff et al., 2004) and on the *CORIS* web interface. A total dataset of 120 concordances per language was extracted. The concordances are normalised for length, measuring between 100 and 200 characters, and averaging around 30 tokens per sentence (English, mean tokens per sentence: 28.67, SD: 2.5; Italian, mean tokens per sentence: 28.65, SD: 5.5).

The concordances were run through the online demo version of Profiling-UD[2]. A total of 14 parameters was selected from all levels of analysis provided by the tool: Raw Text Properties, Lexical Variety, Morphosyntactic information, Verbal Predicate Structure, Global and Local Parse Tree Structures, and Use of Subordination. Table 3 outlines the parameters extracted and their function. The explanation is taken directly from the "Linguistic Profile Legend" provided by Profiling-UD.

Level of analysis	**Linguistic parameters**	**Explanation**
Raw text properties	n_tokens	Total number of tokens
	char_per_tok	average number of characters per word (excluded punctuation)
Lexical variety	Ttr_lemma_chunks100	Type/Token Ratio (TTR) calculated with respect to the lemmata in first 100 tokens of a document. It ranges between 1 (high lexical variety) and 0 (low lexical variety).
	Ttr_form_chunks100	Type/Token Ratio (TTR) calculated with respect to the word forms in first 100 tokens of a document. It ranges between 1 (high lexical variety) and 0 (low lexical variety).
Morphosyntactic information	lexical_density	Ratio between content words (nouns, proper nouns, verbs, adjectives, adverbs) over the total number of words in a document
Verbal Predicate Structure	verbal_head_per_sent	average distribution of verbal heads in the document, out of the total of heads.
	avg_verb_edges	verbal arity, calculated as the average number of instantiated dependency links (covering both arguments and modifiers) sharing the same verbal head, excluding punctuation and auxiliaries bearing the syntactic role of copula according to the UD scheme
Global and Local Parse Tree Structures	avg_max_depth	mean of the maximum tree depths extracted from each sentence of a document. The maximum depth is

[2] The tool is available online at: http://linguisticprofiling.italianlp.it

		calculated as the longest path (in terms of occurring dependency links) from the root of the dependency tree to some leaf.
	avg_token_per_clause	average clause length, calculated in terms of the average number of tokens per clause, where a clause is defined as the ratio between the number of tokens in a sentence and the number of either verbal or copular head.
	avg_max_links_len	mean of the longest dependency links extracted from each sentence of a document.
	avg_links_len	average number of words occurring linearly between each syntactic head and its dependent (excluding punctuation dependencies).
Use of Subordination	principal_proposition_dist	distribution of principal clauses
	subordinate_proposition_dist	distribution of subordinate clauses, as defined in the UD scheme: https://universaldependencies.org/u/overview/complex-syntax.html#subordination.
	avg_subordinate_chain_len	average length of subordinate chains, where a subordinate 'chain' is calculated as the number of subordinate clauses embedded on a first subordinate clause.

Table 3 – *linguistic parameters extracted from Profiling-UD*

After extracting these measures for all the concordances, the dataset was analysed using Principal Component Analysis (henceforth: PCA). Although genre and register analysis generally employs Factor Analysis, it has been shown in the literature that PCA yields results comparable to Factor Analysis, especially if the variables are correlated among them and the number of variables is sufficiently large (Field et al. 2012: 760; Levshina, 2015). The PCA analysis was conducted using the statistical environment R and the package *factoextra* (Kassambara, Mundt, 2020). As figure 1 shows, variables appear to be overall moderately correlated to each other (darker squares indicate higher correlation), justifying the use of PCA as a method of analysis.

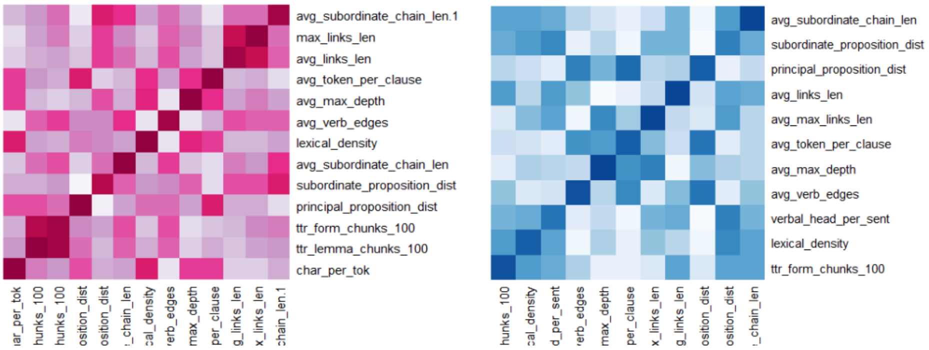

Figure 1 – *heatmaps showing correlations among variables in English (red, left) and Italian (blue, right)*

3.2. Results: PCA of linguistic complexity in Italian LLL

Since the variables extracted from Profiling-UD pertain to both lexical and morphosyntactic level of analyses, we perform two separate PCA analyses: one with more lexically oriented variables (number of tokens, characters per token, lexical density, and TTRs) and on for more syntactically oriented variables (the rest of the 14 variables in table 3).

The scree plots in figure 2 show the percentage of variance accounted for the dimensions for the two analyses. Following the so-called Kaiser criterion, we retain in our analysis only the components with eigenvalues higher than 1(Levshina, 2015:355). Therefore, only the first two dimensions are retained in the analysis of both lexical and syntactic complexity.

Manually inspecting the individual contributions of variables to each dimension (see figures 3 and 4 below) the dimensions of lexical complexity are labelled as (1) *Lexical Diversity*, and (2) *Lexical Density*. For syntactic complexity (figure 4), the dimensions are labelled as (1) *Dependency structure*, and (2) *Internal constituents' complexity*.

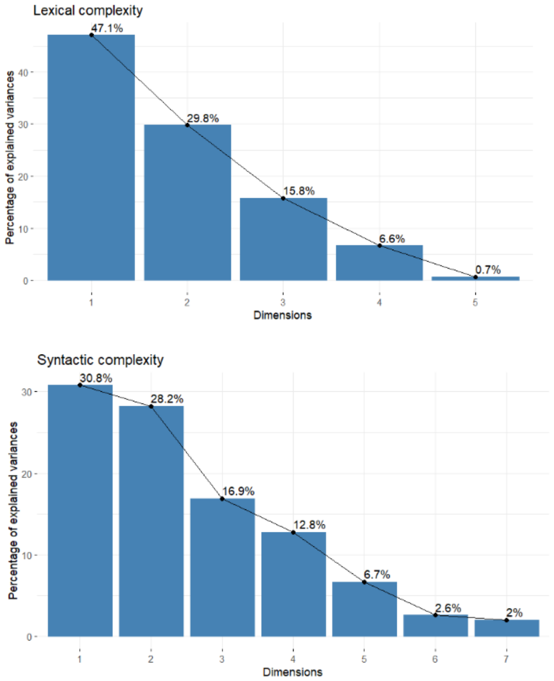

Figure 2 – *scree plot of the percentage of variance accounted for by the different dimensions of lexical and syntactical complexity*

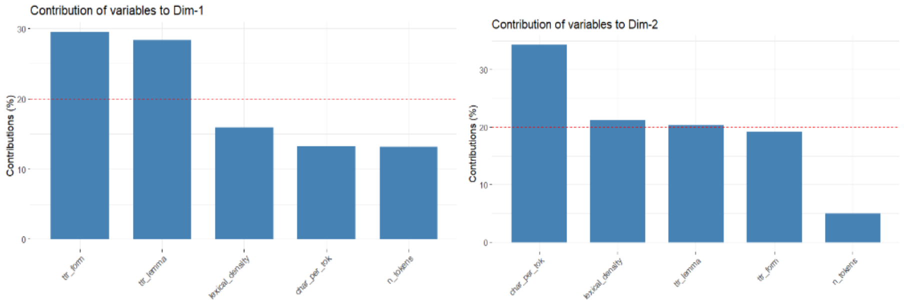

Figure 3 – *individual contributions of variables to the first 2 dimensions of lexical complexity*

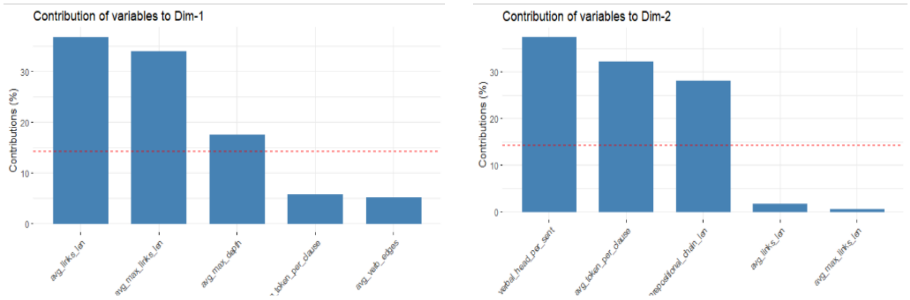

Figure 4 – *individual contributions of variables to the first 2 dimensions of syntactic complexity*

After having identified the principal components of variance, the similarities and differences across the three corpora are explored. As it can be seen from figures 5A and 5B below, lexically it appears that LLL is pretty distant from both legal jargon and written prose. Syntactically – however – LLL appears to be a 'perfect mix' of the two neighbour genres. In other words, it appears that while for morphosyntactic dimensions of complexity LLL is a mix between specialistic and general-domain written language, the lexical aspect of it might be more idiosyncratic.[3]

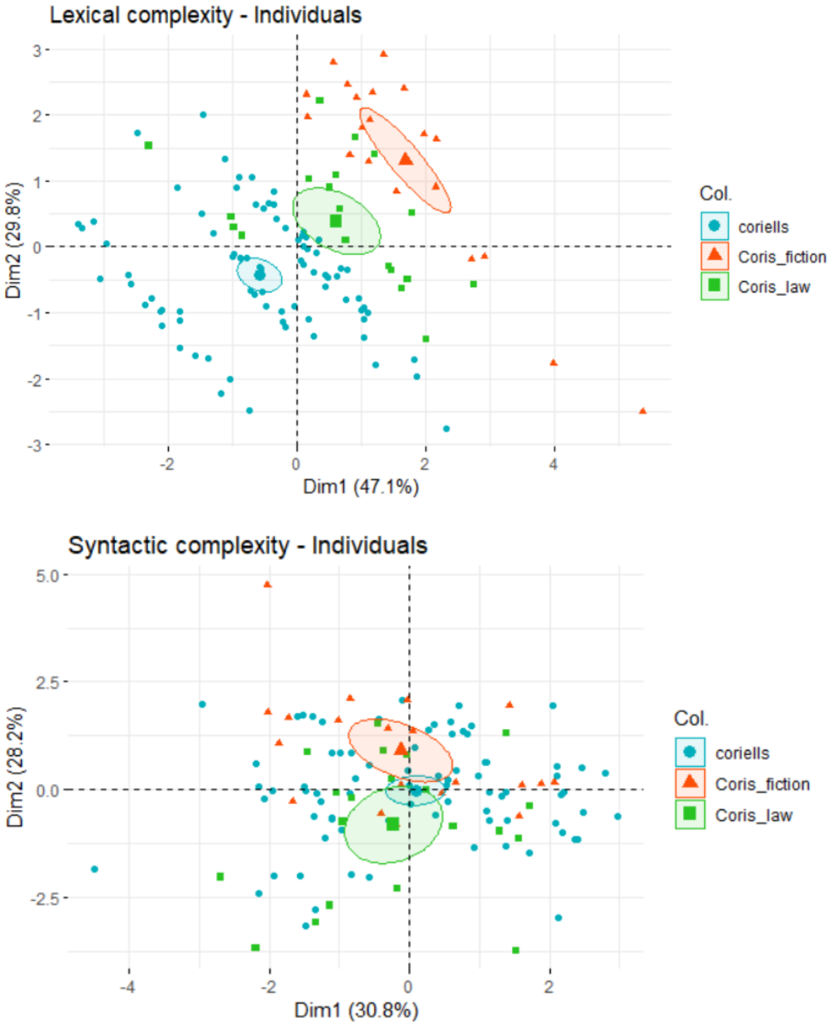

Figure 5 – *first two dimensions of lexical and syntactic complexity for the three corpora. Ellipses represent 95% confidence intervals around the centroid*

[3] It must be noted that the sample of variables that can be extracted from Profiling-UD is unbalanced, with fewer lexical information and the majority of variables pertaining to syntactic profile. However, this tool was used as it is the only one – to the author's best knowledge – that allows for cross-linguistic comparisons for its use of UD.

3.3 Results: PCA of linguistic complexity in English LLL

The same process applied for Italian has been applied to English. The scree plots in figure 6 shows that in this case the first 2 dimensions are considered for lexical complexity, and the first 3 for syntactic complexity. Very similarly to Italian, the dimensions of lexical complexity (see figure 7 below) are labelled as (1) *Lexical Diversity*, and (2) *Lexical Density*. For syntactic complexity (figure 8), the dimensions are labelled as (1) *Syntactic Structure*, (2) *Dependency Length,* and (3) *Internal Constituents' Complexity*.

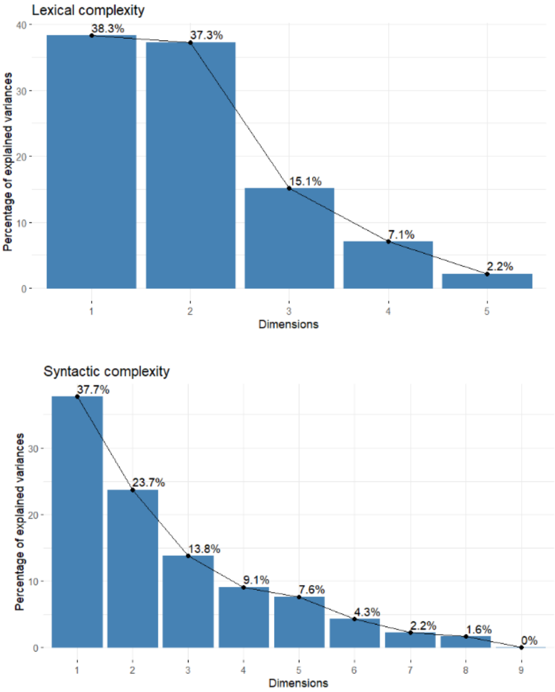

Figure 6 – *scree plots of the percentage of variance accounted for by the different dimensions pf lexical and syntactical complexity*

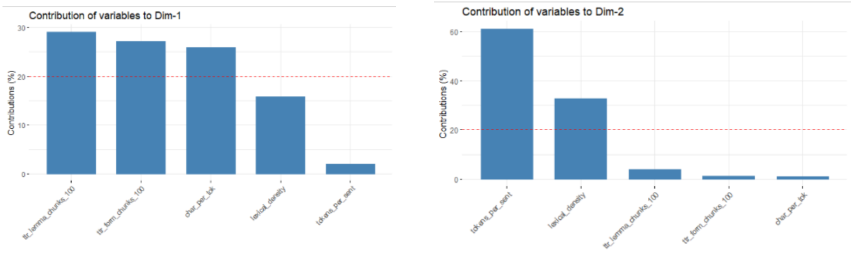

Figure 7 – *individual contributions of variables to the first 2 dimensions of lexical complexity*

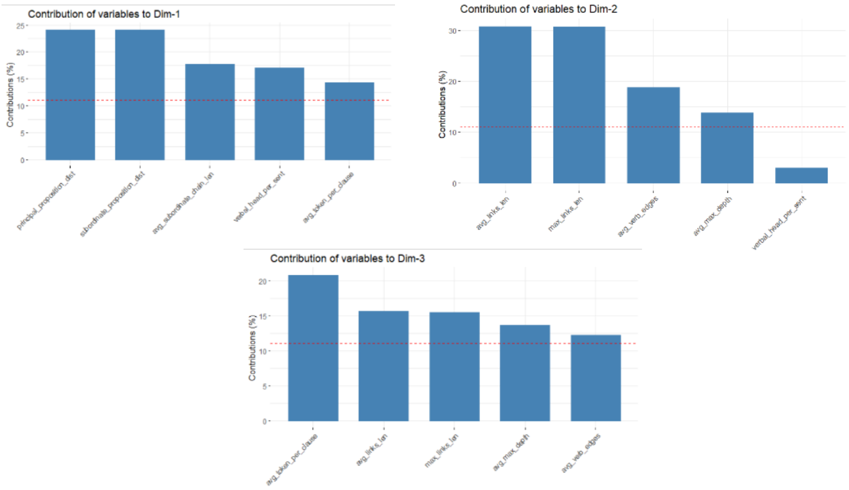

Figure 8 – *individual contributions of variables to the first 3 dimensions of syntactic complexity*

From the distribution of the data across the dimensions of complexity, the situation is quite different from Italian. Lexically (figure 9), English LLL is extremely similar to legal jargon – differently from what was noted for Italian (see section 3.2.1 above). For the three dimensions of syntactic complexity on the other hand there is much more overlap across the three corpora. Particularly, what the graphical overlaps shows is that at the morpho-syntactical level LLL is still similar to legal jargon but with a great influence of general-domain language.

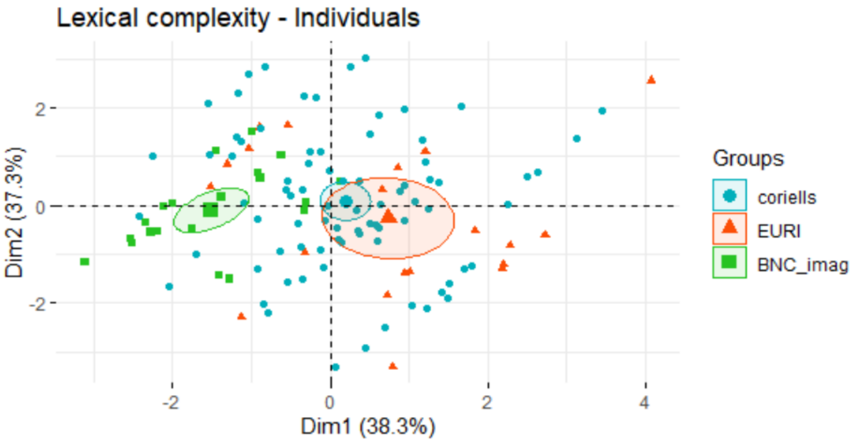

Figure 9 –– *first two dimensions of lexical complexity for the three corpora. Ellipses represent 95% confidence intervals around the centroid*

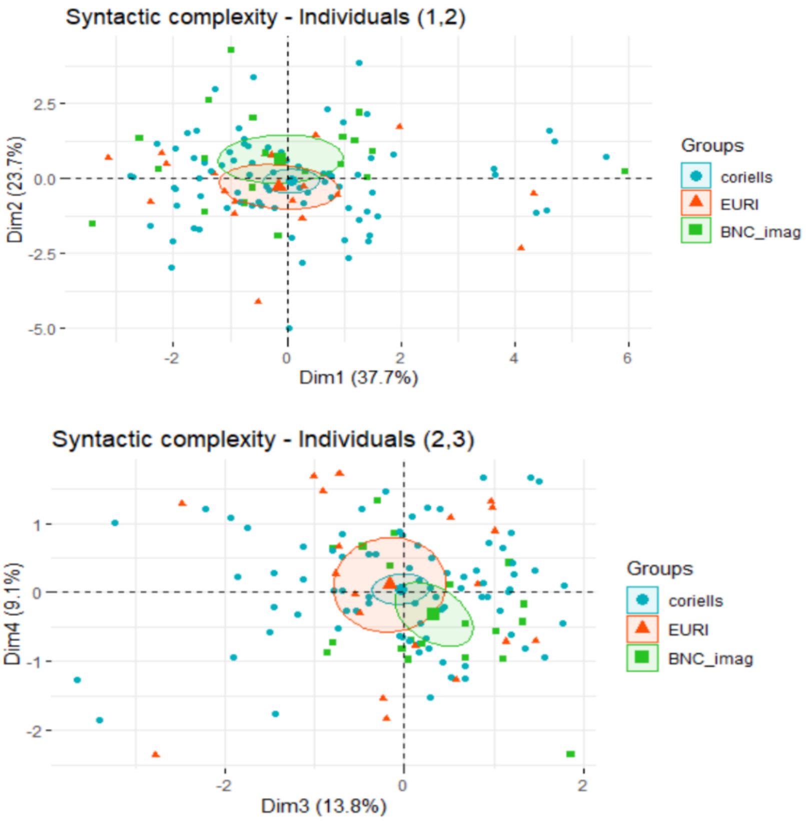

Figure 10 – *first two dimensions of lexical complexity for the three corpora. Ellipses represent 95% confidence intervals around the centroid*

4. *Discussion*

The present paper presented CorIELLS, the first corpus of Italian and English legal-lay language. To show the potential of this corpus and in general of small, specialised corpora, a number of exploratory analyses have been presented. Previous research on CorIELLS mainly aimed at establishing legal-lay language as an idiosyncratic genre, or textual type, based on quantifiable linguistic properties. In this contribution, we expand this line of research by presenting a Principal Component analysis of lexical and syntactical complexity measures.
This type of analysis shows how legal-lay language is positioned across different dimensions of linguistic complexity compared to legal jargon and general-domain prose. Findings, albeit preliminary, seem to provide cross-linguistic support to

our hypothesis of legal lay language being an independent genre. For Italian, we find that along the dimensions of lexical complexity legal-lay language is different from both legal jargon and written prose, while syntactically it appears to be a 'perfect mix' of the two neighbour genres, with vast overlaps. English instead is lexically very similar to legal jargon, while the dimensions of syntactic complexity show much more overlap among the three genres.

To fully understand the differences in the two languages additional research is required, but it appears that English and Italian approach simplification in legal language from two different standpoint: while both English and Italian simplify the syntactic structure (as shown by the overlaps with general written language), at the lexical level English remains quite "conservative", while Italian legal-lay lexicon – while being still closer to legal jargon than to general written language – shows more idiosyncratic characters. Overall, legal-lay language for both Italian and English is confirmed to have structurally autonomous characters, and to be a 'blended' genre made of a mixture of highly specialised features with more general features.

The research outlined here is by no means exhaustive, and possible other lines of research expanding it include a more in-depth look at processing difficulties of legal-lay texts (following for example Gunnarson 1984; Hotta, Fujita 2007; Conklin et al., 2019), especially for vulnerable groups such as L2 speakers or not completely scholarised adults.

Bibliography

Adler, M. (2012). *The Plain Language Movement.* Oxford: Oxford University Press.

Baisa, V., Michelfeit, J., Medveď, M., & Jakubíček, M. (2016). European union language resources in sketch engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 23-28 May 2016, Portorož, Slovenia. European Language Resources Associaton, 2799-2803.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),* May 26-28, 2004, Lisbon, Portugal. European Language Resources Associaton, 1-4.

Benoit, C. (1990). An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity. In *Speech Communication*, 9(4), 293-304.

Benoliel, U. & Becher, S. I. (2019). The Duty to Read the Unreadable. In *SSRN Electronic Journal.*

Bhatia, V. K. (1983). Simplification v. Easification— The Case of Legal Texts. In *Applied Linguistics*, 4(1), 42–54.

Bhatia. V. K. (2004). *Worlds of Written Discourse. A Genre-Based View*. London: Continuum International.

Biber, D. (1993). Representativeness in corpus design. In *Literary and linguistic computing*, *8*(4), 243-257.

Biel, Ł. (2009). Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In *Reconceptualizing LSP. Online proceedings of the XVII European LSP symposium*, 1- 15.

Biel, Ł. (2018). Corpora in Institutional Legal Translation: small steps and the big picture. In Prieto Ramos, F. (Ed.) *Institutional Translation for International Governance. Enhancing Quality in Multilingual Legal Communication*. London/New York: Bloomsbury, 25-36.

BNC Consortium (2007). *The British National Corpus, XML Edition*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554.

Brezina, V. & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. In *Applied Linguistics*, 36(1), 1-22.

Brunato, D. & Venturi, G. (2014). Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici. In *Informatica e diritto*, XL (XXIII), 111-142.

Brunato D., Cimino A., Dell'Orletta F., Montemagni S., & Venturi G. (2020). Profiling-UD: a Tool for Linguistic Profiling of Texts. In *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020*), May 11-16, Marseille, France. European Language Resources Association, 7145–7151.

Busso, L. (2022). Lexicon and Grammar in Legal-Lay Language: a Quantitative Corpus Study in Italian. *Studi Italiani di Linguistica Teorica e Applicata,* LI(1), 5-32.

Busso, L. (forthcoming) An investigation of the lexico-grammatical profile of English legal-lay language. *Journal of Language and Law.*

Chovanec, J. (2013). Grammar in the Law. In Chappelle, C. (Ed.) *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell, , 1- 8.

Conklin, K., Hyde, R., & Parente, F. (2019). Assessing plain and intelligible language in the Consumer Rights Act: a role for reading scores? In *Legal Studies*. 39(3), 378-397.

Crystal, D. & Davy, D. (1969). *Investigating English style*. London, Longman.

Curtotti, M. & McCreath, E. (2013). A right to access implies a right to know: An open online platform for research on the readability of law. In *Journal of Open Access to Law* 1(1), 1-56.

De Mauro, T. & Chiari, I. (2016). *Il Nuovo vocabolario di base della lingua italiana*, available at: https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base- della-lingua-italiana.

Eagleson, R. (2004). Plain Language.Gov- Improving communication from the Federal government to the public. Retrieved on January 10, 2022 at http://www.plainlanguage.gov/whatisPL/definitions/eagleson.cfm.

Fairclough, N. (2011). *Critical discourse analysis- The critical study of language. 2nd ed.* Harlow: Longman.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using* R. Los Angeles: Sage.

Flowerdew, L. (2002). Corpus-based Analyses in EAP'. In Flowerdew, J. (Ed.) *Academic Discourse*. London: Pearson, 95–114.

Flowerdew, L. (2004) The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings. In Connor, U. & Upton, T. (Eds) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11–33.

Frade, C. (2007). Power dynamics and legal English. In *World Englishes*, *26*(1), 48-61.

Frade, C. (2016). The power of legal conditionals in international contracts. In Hafner, C.A., Wagner, A., Bhatia, V.K. (eds.) *Transparency, Power, and Control*. London: Routledge, 45-64.

Garzone G. (2008). International commercial arbitration rules as translated/re-written texts: an intercultural perspective. In Bhatia, V. K., Candlin, C. N. & Evangelisti Allori, P. (Eds.) *Language, Culture and the Law. The Formulation of Legal Concepts across Systems and Cultures*, Frankfurt am Main: Peter Lang, 47-73.

Gibbons, J. (2003). *Forensic linguistics: An introduction to language in the justice system*. Oxford: Blackwell.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A. E. (2019). *Explain Me This*. Princeton: Princeton University Press.

Gunnarsson, B.-L. (1984). Functional comprehensibility of legislative texts: Experiments with a Swedish act of parliament. In *Text - Interdisciplinary Journal for the Study of Discourse*, 4(1–3), 71-106.

Hotta, S. & Fujita, M. (2012). The psycholinguistic basis of distinctiveness in trademark law. In Tiersma, P. M., & Solan, L. M. (Eds.) *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press.

Hyland, K. (2006). *English for Academic Purposes: An advanced resource book*. London: Routledge.

Kassambara, A. & Mundt, F. (2020). *Factoextra Extract and Visualize the Results of Multivariate Data Analyses*. R Package Version 1.0.7. https://CRAN.R-project.org/package=factoextra

Kilgarriff, A., P. Rychlý, P. Smrz, & D. Tugwell. (2004). Itri-04-08 The Sketch Engine. In *Information Technology,* 105- 116.

Kimble, J. (2000). The Great Myth That Plain Language Is Not Precise. In *Scribes Journal of Legal Writing,* 7*,* 109-118.

Koester, A. (2010). Building small specialised corpora. In O'Keeffe, A., & McCarthy, M. (Eds.). *The Routledge handbook of corpus linguistics*. London: Routledge, 66-79.

Kusters, W. (2003). *Linguistic complexity: The influence of social change on verbal inflection*. Leiden: LOT.

Kusters, W. (2008). Complexity in linguistic theory, language learning and language change. In Miestamo, M., Sinnemäki, K. & Karlsson, F. (Eds.) *Language complexity: Typology, contact, change.* Amsterdam: John Benjamins, 3-22.

Levshina, N. (2015). How to do linguistics with R. *Data Exploration and Statistical Analysis.* Amsterdam-Philadelphia: John Benjamins.

Matulewska, A. (2007). *Lingua legis in translation*. Frankfurt am Mein: Peter Lang.

Mori, L. (2019). Complessità sintattica e leggibilità. Un monitoraggio linguistico per la valutazione dell'accessibilità dei testi legislativi europei e italiani. In *Studi Italiani di Linguistica Teorica e Applicata*, 48, 627-657.

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

Nencioni G. (1983). Parlato-parlato, parlato-scritto, parlato-recitato. In Nencioni G., (ed.), *Di scritto e di parlato. Discorsi linguistici* (pp. 126-1790, Bologna: Zanichelli.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, 3–16.

Pallotti, G. (2015). A simple view of linguistic complexity. In *Second Language Research*, 31(1), 117-134.

Rossini-Favretti, R., Tamburini, F. & Desantis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Wilson, A., Rayson, P. & Mcenery, T. (Eds.) *A rainbow of corpora: Corpus linguistics and the Languages of the world.* Munich: Lincom-Europa, 27-38.

Petyko, M., Atkins, S., Busso, L. & Grant, T. (forthcoming). The Forensic Linguistic Databank. *Journal of Language and Law.*

Stefanowitsch, A. (2013). Collostructional analysis. In Hoffmann, T. & Trousdale, G. (Eds.) *The Oxford handbook of Construction Grammar*. Oxford: Oxford University Press.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tiersma, P. M. (1999). *Legal language*. Chicago: University of Chicago Press.

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Van Boom, W. H., Desmet, P., & Van Dam, M. (2016). "If It's Easy to Read, It's Easy to Claim"—The Effect of the Readability of Insurance Contracts on Consumer Expectations and Conflict Behaviour. In *Journal of Consumer Policy*, 39(2), 187–197.

Venturi G. (2011), Semantic annotation of Italian legal texts: a FrameNet-based approach. In *Constructions and Frames* 3(1), 46-79.

Williams, C. (2004). Legal English and Plain Language: an introduction, In *ESP Across Cultures* (1), 111-124.

Williams C. (2004b). Pragmatic and cross-cultural considerations in translating verbal constructions in prescriptive legal texts in English and Italian. In *TEXTUS* XVII/1, 217-246.

Williams, C. (2005). Progetto Chiaro! and the Plain Language Movement in Italy. In *Clarity* 53, 30-32.