UNIVERSITY OF SIENA
DEPARTMENT OF SOCIAL, POLITICAL AND COGNITIVE SCIENCES

DOCTORAL PROGRAM IN COGNITIVE SCIENCES

XXVI° CYCLE

A STUDY ON LINGUISTIC COMPLEXITY FROM A COMPUTATIONAL LINGUISTICS PERSPECTIVE. A CORPUS-BASED INVESTIGATION OF ITALIAN BUREAUCRATIC TEXTS.

MAJOR IN LINGUISTICS (L-LIN/01)

Author: Dominique Brunato

Advisor: Prof. Adriana Belletti

A.A.: 2014/2015

ACKNOWLEDGMENTS

I wish to express my gratitude to my supervisor, Professor Adriana Belletti, who first introduced me to linguistics research and let me the freedom to explore a topic which was quite different from what initially expected, always reminding me to dig beyond the surface of linguistic data for a deeper understanding of language-related phenomena.

I am also extremely grateful to all the members of the ItaliaNLP Lab at the Institute of Computational Linguistics "Antonio Zampolli" of CNR in Pisa, for giving me the opportunity to take part in their stimulating research. I especially thank Giulia Venturi, Felice Dell'Orletta and Simonetta Montemagni for their suggestions, feedback and original contributions in developing many of the ideas contained in this thesis.

A special thought goes to all the young and enthusiastic researchers I met during these years of my PhD, at Ciscl (Centro Interdipartimentale di Studi Cognitivi sul Linguaggio), the University of Siena, the UCREL Summer School at Lancaster University, in particular David Cioncoloni, Giulio Mecacci, Daniele Botteri, Stefano Piazza, Candice Coyer, Maggie Leung, Ivana Lalli Paćelat, Marek Łukasik, Kathi Franko, and I wish them to always find passion in their research.

An important step of my training in the field of computational linguistics was the period I spent in the company Expert System in Modena, and I want to thank all the very brilliant minds I met there.

Just for being always by my side, a lovely thank to my mother, my sister Marika, Enrico and Rachele, the most precious gift I got during my thesis.

Table of Contents

Chapter 3

Chapter 4

I. Abstract

This thesis investigates the construct of linguistic complexity from a *user-based* perspective and its computational treatment from the applicative viewpoint of automatic readability assessment of written texts and, specifically, of Italian bureaucratic texts. Such a choice has been motivated by the well-known complexity of bureaucratic language, the so-called "bureaucratese"[1], which tends to be unnecessarily distant from the variety of standard language and to resemble instead the language of the law, despite the different intended audience.

A feasible way to enhance the comprehension of bureaucratic texts relies today on the use of advanced language technologies, and particularly those devoted to assessing the readability level of a text. This has also been foreseen by the more recent publication on the simplification of Italian bureaucratic language, which encourages public employees, when faced with the task of making their documents more comprehensible, to «build sentences taking into account the limits to readability in according to current indexes»[2].

But can current indexes – and particularly those available for the Italian language – discriminate between unnecessary complexity, namely typical *bureaucratese* markers, and other "genre-specific" complexity features?

This general question can be decomposed into finer sub-questions, with both theoretical and applicative implications for the study of linguistic complexity and its computational treatment, such as:

- Which features of a text embody a general, i.e. valid "across textual genres", notion of linguistic complexity?
- Which features of a text embody a "genre-specific" notion of linguistic complexity, such as the one characterizing the domain of bureaucratic language?

[1] Terms such as "bureaucratese", "officialese" or "gobbledygook" all designate bureaucratic language with a negative nuance.

[2] We refer to the *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*, edited by the Italian Institute of Legal Information Theory and Techniques (Istituto di Teorie e Tecniche dell'Informazione Giuridica – ITTIG) and the Accademia della Crusca in 2011. This is the outcome of a project carried out by a working-group composed by linguists, lawyers and public officials, with the specific aim of improving the comprehensibility and communication capacity of administrative acts.

- Is this twofold typology of features already handled by current readability assessment indexes? And, if the answer is no, how can an automatic system succeed in learning the difference?

- In what way linguistic complexity features for readability assessment can make it possible the automation of related and, more specific applicative tasks, such as text simplification?

This thesis aims at shedding light on these questions and it is organized into three main parts, which will be better detailed in section II.

A **first part** (chapters 1-2) provides a critical overview of the literature on the construct of linguistic complexity from a *user-based* viewpoint, which is addressed primary from the perspective of formal and empirical studies on language within the domain of cognitive sciences, and then by focusing on the computational linguistics perspective as it has been currently operationalized in automatic readability assessment research.

The **second part** (chapter 3) describes an original Natural Language Processing (NLP)-based investigation carried out on a corpus of Italian bureaucratic texts through the methodology of linguistic profiling. The intent has been precisely to characterize the notion of linguistic complexity within this domain, with a view to the specialization of a "general-purpose" readability index towards this textual genre.

The **third part** (chapter 4) broadens the view to the applicative scenarios of the computational measurement of linguistic complexity, by focusing on the field that is more directly related to readability assessment research: Automatic Text Simplification (ATS). More specifically, it will be here presented the outcomes of an ongoing study for the Italian language, which has led to the development of a new annotation scheme for the analysis of linguistic phenomena involved in manual text simplification, as a preliminary step for a semi-automatic and automatic treatment.

In the **conclusion** (chapter 5) we summarize the main findings of the whole work, as well as some promising research perspectives which, from multiple viewpoints, are all concerned with the investigation and applicative treatment of linguistic complexity in texts.

II. Thesis overview

The main sections sketched above are organized into the following four main chapters:

Chapter 1 reflects upon the topic of linguistic complexity in natural language data trying to elucidate how it has come to be characterized by current linguistic research within the framework of cognitive science. It thus takes into account a set of properties of linguistic units, spanned across the domains of words, phrases, sentences and discourse, which have proven to modulate the difficulty of human sentence processing. A particular interest will be devoted to deepening the notion of syntactic complexity: this is indeed the domain in which the interplay between the empirical methods of psycholinguistics and the analytical tools developed by formal syntax has been particularly fruitful, contributing to provide a more comprehensive view of the nature and processes underlying sentence reading and comprehension.

Chapter 2 addresses text difficulty analysis from the computational linguistics perspective; more in detail, it focuses on the operationalization of linguistic complexity to enable applicative research in automatic readability assessment of texts. Taking as a point of the departure a definition of the concept of readability, in which it will be pointed out its controversial use to designate interrelated but not interchangeable notions – such as "understanding", "ease of reading", "comprehensibility" – paragraph 2.2 reviews the former approach to readability assessment based on traditional readability formulae. Despite their ongoing popularity, these formulae seem nowadays too simplistic with respect to both the theoretical ground and the methods they adopt to estimate the difficulty of a text. On the other side, it will be highlighted how the progress in language technologies and statistical methods for text analysis has allowed for a more fine-grained, as well as cognitively aware, treatment of linguistic complexity proxies within texts: this is the rationale behind the current "second-generation" readability indexes, which will be covered in section 2.3. In particular, a more in depth presentation will be dedicated to READ-IT (§2.4), which is the first NLP-based tool specifically designed to assess the readability level of Italian texts. As we will see, a qualifying feature of READ-IT is to be a "general-purpose" readability assessment system, thus trained on corpora representative of standard Italian. However, it is not necessarily true that the variety of standard (Italian) language displays the same peculiarities of more complex language varieties, such as the bureaucratic language.

The evaluation of the influence of textual genre on the automatic assessment of readability has provided the leading research question around which **Chapter 3** develops. In this chapter, a "quasi-parallel corpus" of Italian bureaucratic texts has undergone a comparative *linguistic profiling* investigation by exploiting NLP-enabled features, with the aim of evaluating whether and to what extent the peculiarities of bureaucratic language might impact on standard readability assessment models. The theoretical and operative requirements of the linguistic profiling methodology, as well as the design and pre-processing of the corpora, will be covered in **paragraphs 3.3-3.5**, while section **3.6** will illustrate the resulting quantitative findings. Although significant correlations with traditional descriptive studies on Italian bureaucratic language and its deviant patterns from standard language have been found, in **section 3.7** we will also take into account some shortcomings of the automatic representation adopted here as the only mean to characterize where syntactic complexity might stem from within written texts. In this context, a qualitative investigation of the corpus, focused on a well-known typology of difficult sentences, i.e. relative clauses, has been carried out and inspired to a featural approach to syntactic locality developed by formal linguistic theory.

Finally, **Chapter 4** is intended to outline the first achievements of an ongoing investigation into Automatic Text Simplification (ATS), which can be properly considered as a natural step forward of readability assessment research. Despite the popularity it is enjoying within the international computational linguistics community, such a topic is relatively un-researched for the Italian language, mainly due to the lack of language-specific resources. This chapter illustrates a study which has tackled the construction of the necessary tools and resources as a preliminary development to inform research on both automatic and semi-automatic text simplification for Italian.

Original contributions

The following major goals have been achieved from this thesis:

i) on a theoretical side, we highlight the fine-grained nature of the construct of linguistic complexity, which derives from the complexity itself of the process of language comprehension as a cognitive capacity. If such a condition makes qualitative linguistic research fundamental to inform any operationalization of text difficulty analysis with respect to the intended user, we also wish to demonstrate that NLP-based technologies for text accessibility are nowadays able to intercept a rich and comprehensive set of linguistic complexity proxies, which are as also cognitively adequate in light of empirical research on human language comprehension. This holds despite it is acknowledged that the accuracy of automatic linguistic annotation, which stands at the core of humane language technologies, tends to decrease at more sophisticated levels of text analysis, e.g. syntax, especially in out-domain scenarios, i.e. when analysing documents not belonging to the domain of training data, such as the bureaucratic texts.

ii) with respect to most current research in automatic readability assessment, it is here presented a case-study which provides evidence that such a kind of task requires a "genre-specific" notion of linguistic complexity, capable of intercepting the linguistic peculiarities of the domain under analysis;

iii) from a methodological perspective, the linguistic profiling investigation discussed in Chapter 3 would like to offer an innovative perspective to enrich the wide literature devoted to investigating the relation between standard Italian language and bureaucratic language, as well as provide measures for enhancing the quality of Italian official writing;

iv) with respect to Automatic Text Simplification, the study described in Chapter 4 introduces an original approach to this topic, whose qualifying aspects have been the design and development of a new annotation scheme, grounded on the linguistic and psycholinguistic literature on text complexity, for the classification of diverse typologies of linguistic operations which have to be expected when analysing manually simplified texts.

Such a scheme has been tested on two sentence-aligned corpora, providing a new resource specifically conceived to explore text simplification for the Italian language from a data-driven approach and with a view to its automatic treatment.

Chapter 1

Cognitive approach to the study of linguistic complexity: what theory and data suggest

1.1 Introduction

The study of complexity of natural language is a central and debated topic of linguistics research, which has been addressed, and it still being addressed, from a multidisciplinary perspective, ranging from typological studies to historical linguistics, from psycholinguistics to language acquisition, from neurosciences to computational linguistics. According to the framework, several attempts have been set forth not only to provide a definition, but also a metric to evaluate and operationalize it. Broadly speaking, this construct can be investigated from two main approaches: the *grammar-based* and the *user-based* approach, also referred to, respectively, as absolute and relative complexity (cf. Blache, 2011).

The former considers linguistic complexity as an inherent property of the system that can be assessed by comparing languages along several formal properties, such as the number of rules to yield a certain output, the number of exceptions to the rules or the size of linguistic inventories at varying levels of representation (e.g. the number of phonemes). Under this perspective, some authors have introduced a hierarchy of complexity, so that «a complex language is one which, if compared to a simpler one, contains more overt signaling of various phonetic, morphological, syntactic and semantic distinctions beyond communicative necessity» (McWorther, 2001:25). On a similar ground, Ferguson and DeBose (1977) led their analysis of pidgin and creole languages, assigning them the status of simplicity because, according to the authors, «they omit material, reduce irregularity, or make sound-meaning correspondences more transparent»; all of these «modifications, intended in a fairly obvious way to make utterances easier to perceive, understand, or produce, may be regarded as simplifying processes».

Despite these positions – as Newmeyer and Preston (2014) point out in their introductory chapter to *Measuring grammatical complexity* – by the late twenties, a large consensus has

been established among linguists, regardless their orientation, in considering all languages equally complex. Conversely, the interest has been focused on deepening the "trade-off hypothesis", according to which, when languages are comparatively investigated along different grammatical components, it appears that a source of complexity within a domain (e.g. a rich case marking), it is balanced by a relative simplicity within another (e.g. a flexible word order).

The second approach to linguistic complexity – which was profoundly influenced by the so-called "Chomskian revolution"[3] – is more subjective and user-dependent and evaluates complexity in terms of processing difficulty. A key intent of this research paradigm is to validate the implications underlying formal theories against the performance evidence, thus taking into account the viewpoint of the user engaged in the process of language understanding. Over the last fifty years, indeed, the investigation into the faculty of language has become prominent across diverse disciplines in the realm of cognitive sciences, all of which have contributed to provide a clearer picture of the process of language comprehension, along with the constraints (in terms of mental architecture, execution and computational resources) it is subject to. While there is no agreed consensus on the functioning of the general processing mechanisms[4], it is well established that language comprehension is a real-time process, which requires the mental parser to access, and

[3] As pointed out in (Searle, 1972), the "revolution" that Chomsky (1957, 1959, 1965) brought in the field of linguistics can be primarily considered as a redefinition of the subject matter itself, which became the faculty of language as an organ of the mind; such a change has led, consequently, to rethink both the scopes and the methods employed by traditional linguistic inquiries based on structuralism. See also Rizzi (1998, 2003) for an overview of the cognitive paradigm applied to linguistic research and its implications for the study of related domains, particularly language development, adult comprehension and language pathologies.

[4] Two major accounts have been put forth to explain sentence processing mechanisms, namely the serial (or dual-stage) and the interactive account, which descend in turn from two diverse approaches to the study of language comprehension: the principle-based approach and the constraint-based approach, respectively (see Pickering and van Gompel, 2006; Harrington, 2001 for a survey). The principles-based approach assumes a modular model of the mind (Fodor, 1983), according to which language is a separate, specialized component that exists independently of a central store of general knowledge. The language module, in turn, is composed by other sub-modules, each one accounting for specific aspects of language processing; among these, the syntactic sub-module is viewed as the only responsible for the initial parsing of incoming word strings, whereas other sources of knowledge (i.e., lexical, pragmatic, real-word) are assumed to become available later. Under this perspective, sentence processing is carried out serially, with the priority of syntax. The Garden-Path theory (Frazier, 1987; Frazier and Fodor, 1978) is certainly one of the most representative models of sentence processing based on the principle-based account.
On the contrary, the constraint-based approach (MacDonald, Pearlmutter and Seidenberg, 1994; Tanenhaus and Trueswell, 1995) lies on a connectionist view of the mind architecture (McClelland, 1988), for which language knowledge is not stored symbolically but rather distributed in associated patterns of neural networks. Within these patterns, all the sources of linguistic knowledge interact simultaneously and constraint the online comprehension from the very beginning. Thus, the processing is interactive and produces all the analyses that are compatible with the ongoing input in cascaded fashion.

progressively integrate in a coherent representation, all information deriving from increasingly higher levels of linguistic domains, in order to derive the meaning of the whole sentence. Besides, since language is part of a dynamic neural architecture, linguistic computations occurring online interact with the neural and cognitive constraints, both language-specific and domain-general; as a consequence, language comprehension can be more demanded to the extent to which the "core" properties underlying linguistic computations contrast with economy processing principles (or "least effort" conditions), in that they require, among others, a bigger storage of working memory resources, additional stages of analysis, some reference to the extra-linguistic knowledge or, when syntax is concerned, a deviation from parser's initial biases.

Particularly fruitful in establishing a "repertoire" of linguistic structures whose fine-grained properties impact on processing effort is the endorsement of behavioural data, in the way they emerge from akin areas of study: among all, adult psycholinguistic performance, language acquisition and language pathologies. With this respect, it is especially in the domain of syntax that the cognitive viewpoint to the study of natural language has contributed to endow the theoretical formalisms with that level of "explanatory adequacy" which, in Chomsky's seminal work (1965), constitutes the highest one that a theory of grammar can meet[5].

Next paragraphs will focus more in depth on this second perspective to linguistic complexity, with the aim of reviewing some principles and factors that have been proposed to explain linguistic complexity with respect to human performance both in normal and language-impaired adults, as well as in children. In discussing them, we will follow the sequential processing stages involved in language comprehension and, especially, in reading processes (cf. Perfetti, 1985), thus distinguishing lexical access, syntactic parsing and the interpretation of sentences within the discourse.

[5] In *Aspects of the Theory of Syntax* (1965), Chomsky introduced a hierarchy of three "levels of adequacy" that a linguistic theory might potentially meet - i.e. observational, descriptive and explanatory - and he considered the third one as proper of a generative theory of grammar. In more recent developments, a fourth level is added, which is «deeper than explanatory adequacy, asking not only *what* the properties of language are, but *why* they are that way» (Chomsky, 2001:2).

1.2 Lexical complexity

The first step in language comprehension consists in the process of lexical access, also called *word parsing* (Seidenberg, 1989: 54), i.e. the mental operation whereby the parser associates the symbolic representation of the word to its memory unit. Current models of mental lexicon assume that words are stored in long-term memory as rich lexical entries. This means that they correspond to clusters of representations making reference to diverse theoretically distinct vocabulary types, which overlap with the vocabulary of syntax, semantics, morphology and phonology.

As a consequence of the mental lexicon structure, word complexity cannot be tackled as a unitary phenomenon but instead as a heterogeneous one, deriving by the internal properties of each subcomponent of the correspondent entry. Broadly speaking, all these properties do not necessary make the corresponding entry more difficult to process, but they seem to affect subject performances in a «post-access decision process»[6], i.e. after lexical access[7].

Let us begin by focusing on the variables that influence the recognition of isolated (written) words. At this level, one of the most robust findings in psycholinguistic research is the *word frequency* effect, which accounts for the fact that low-frequency words tend to be read slower than high-frequency ones (Segui *et al.*, 1982).

However, the negative bias for low-frequency words can be modulated by both extra-linguistic and linguistic factors. The former account for the "attitude" of the speakers/readers towards a specific word, which can be assessed in terms of *age of acquisition* and *subjective familiarity* (Burani et al., 2001, for a review). With respect to the latter, it is necessary to consider also the morphological structure of the words, as it has been proven that the frequency of their internal constituents can work as a further discriminative variable. This effect is called *root frequency effect* and it refers to the fact that derivative low-frequency words are recognized more quickly when they contain a more productive root in the target language (see Laudanna and Voghera, 2006 for a review). On the other side, reading performance for low-frequency words negatively correlates with increased *word length*, and this is especially evident in dyslexic readers (Burani *et al.*, 2001).

[6] Cf. the the review reported in (Cutler, 1983).

[7] Experimental research in language comprehension/production exploits a variety of methodological paradigms to collect performance data; for what concerns behavioral tasks - i.e. tasks where the subject is presented with a linguistic stimuli and asked to perform an action (e.g. take a lexical decision, provide a grammaticality judgment or just reading it) - higher response latency, reading times or error-rates are generally taken as a measure of processing difficulty.

When the orthographic representation of the word is taken into account, there is has another property capable of modulating lexical activation and its related processes: the so-called *orthographic neighbourhood*. As originally described in Coltheart *et al.* (1977), orthographic neighbours are those words that can be derived by changing one letter of the stimulus item, preserving letter positions[8]. Two related effects of the orthographic neighbourhood have been mainly investigated: the neighbourhood size or density (also called N-size) and the neighbourhood frequency. What emerges from empirical tasks (cf. Perea and Rosa, 2000 for a comprehensive summary) is that, while the presence of a wider N-size tends to elicit lower reading times and a general facilitative effect for the target word - which is greater, again, for low-frequency words - the effect of having high frequency neighbours is reversed, i.e. inhibitory. Interestingly, the advantage of a dense N-size on low-frequency words has been recently assessed in children with dyslexia[9] too, thus proving the appropriateness of this variable to serve as a proxy of lexical complexity, as well as its effectiveness, in combination with other factors, to reduce the burden of word recognition, which is typically encountered by this population.

Lexical activation also undergoes the effects of variables taking into account conceptual properties of lexical meaning. For instance, *concreteness*, defined as «the property of words referring to objects, animate beings, actions, or materials that can be experienced directly by the senses» (Paivio *et al.*, 1968; Roncato, 1974), has been found to positively affect lexical decision times (Bleasdale, 1987), and again for low-frequency words (James, 1975)[10]. Yet, if abstract words tend to affect reading performance, concrete words, in their turn, can be more or less complex to process according to the influence of other properties. It is worth mentioning here the well-studied *prototypicality effect*, which stems from the property of a word to denote the prototypical member of the corresponding conceptual category.

Under the psychological model outlined by the *Prototype Theory* (Rosch, 1976), the process of categorization starts from the identification of the most typical member of a cognitive category (i.e. the prototype). After this stage, all other entities sharing a higher number of features with the prototype are aggregated around it, so that a category is cognitively established.

[8] For instance, the Italian adjective "solare" (sunny) has the noun "molare" (molar tooth), and the verbs "colare" (to pour slowly, to strain) and "volare" (to fly), among its neighbours.
[9] Cf. data reported by a recent study of Marinelli *et al.* (2013) on Italian speaking children.
[10] Quotations reported in Barca *et al.* (2001).

More specifically, this model assumes that concepts are stored in mind in terms of sets of attributes, which are arranged into a three-leveled hierarchy. The first level is composed by a "core" subset of features, i.e. the features maximizing the distinctiveness of the category against the others, and they correspond to the features embodied by the prototypical members. Other two peripheral subsets denote the superordinate- and subordinate-leveled features: the former provide an abstract characterization of the concept, since they identify those properties that overlap among the categories whose members are internally further distinguishable; the latter, instead, minimize the distinctiveness of the category, as they did not add any further information that is not already instantiated by the prototype. A well-known example takes into account the concept of BIRD[11], whose prototypical member is more likely to be a ROBIN rather than a PENGUIN, since the former shares a greater set of features denoting a prototypical bird than the latter.

For the purposes of characterizing lexical complexity, it is worth noting that the *Prototype Theory* has also crucial implications for the mental lexicon organization, since it suggests that the way we build conceptual categories affects word storage and retrieval. In particular, it has been empirically assessed that nouns denoting the prototypical member (*i.e.* basic level noun) prime faster recognition of the underlying concept than nouns denoting non-prototypical members, although they were both judged as being related to the superordinate concept by the same taxonomic relation (Rosch, 1976).

When the role of semantics in word processing is considered, the empirical findings seem to be more controversial. In this regard, it is necessary to point out that the concept of lexical semantic complexity itself may cover multiple aspects of the meaning of a word. Not only it refers to *lexical ambiguity* (which is one of the "signatures" of the more pervasive phenomenon of language ambiguity), but also to the effects deriving by more abstract semantic properties of lexical items, such as lexical selection restrictions, presuppositions, implications and coercion mechanisms[12], which all contribute to limit the type of contexts a word can stay in.

[11] Capitalization is used to refer to the concept.

[12] In compositional semantics analyses, coercion is a kind of semantic typing adjustment which accounts for the mismatches between selecting and selected type (see, Pustejovsky, 1995). For instance, when a predicate like *begin* combines with a noun phrase describing an entity, as in *begin the book*, the noun phrase comes to assume a silent event interpretation, such as *begin <u>reading</u> the book*, thus undergoing a "type shifting" rule. (Partee and Rooth, 1983).

Let us consider lexical ambiguity, which is distinguishable, in turn, into unsystematic (i.e. when the same word denotes multiple but unrelated referents) and systematic (i.e. when the same word bears related senses in different categories, thus a case of syntactic ambiguity).

With respect to unsystematic lexical ambiguity, psycholinguistic evidence seems to rule out ambiguous words as a source of greater processing load. Indeed, in typical lexical decision tasks (i.e., deciding if a visually presented string of letters is a word or a nonword), subjects have demonstrated to be equally fast in recognizing an ambiguous word, whether the previous sentence primed or not the compatible reading of it. Ambiguity plays a role in a post-lexical access phase, i.e. when the speakers have to elaborate an interpretation of the sentence as a whole: such an effect is compatible with data showing higher reaction times when readers are asked to judge the acceptability of sentences containing ambiguous words.

However, knowing a word also means knowing the grammatical category it belongs to, the number of arguments it requires (i.e. the *theta-grid* associated with), as well as the way they are syntactically realized. In the next paragraph we will focus more in detail on these variables by outlining the nature of the effects they raise in processing.

1.2.1 Lexicon-syntax mapping

The paradigmatic case for analysing the role of the mapping principles between lexical semantic and syntactic information is offered by verbs, whose great variety of argument structures has proven to influence the ease with which incoming words come to be incorporated into higher-level syntactic objects (phrases and sentences), especially for particular populations of "atypical" speakers. This occurs in at least two senses accounting for: i) the number and typology of arguments selected by the verbal entry and ii) the effect of the thematic hierarchy, which establishes how semantic roles will be mapped onto syntactic arguments.

A) **The number and typology of arguments selected by the verbal entry**

With respect to the first issue, it has been shown that verbs selecting only one argument (e.g. the external argument for unergative verb or the internal arguments for unaccusative verbs) are processed faster than verbs selecting two or more obligatory arguments (Shapiro *et*

al., 1987). This holds particularly for agrammatic Broca's aphasics, who have more difficulty in retrieving verbs requiring more obligatory arguments (Thompson *et al.*, 1997).

However, following a seminal work of Fodor, Garrett and Bever (1974), a great deal of psycholinguistic studies has proven that what impacts more on verb processing is not the simple number of arguments but the number of potential argument structures associated with a verb, so that the richer is the number of subcategorization frames in which a verb appears, the longer it will take to process a sentence containing that verb. This implies that, e.g., a verb like 'studiare' [to study], which allows both the transitive structure (e.g. *Rachele sta studiando matematica* [Rachel is studying maths]) and the unergative structure (e.g. *Rachele sta studiando* [Rachel is studying]) is more demanding than a verb like 'dormire' [to sleep], which allows only the second structure.

If this property has a general impact on processing, it is also the case that verbs with a similar argument structure pose different demands on the interpretation of the sentences in which they occur. Let's focus on a particular class of verbs, i.e. those selecting as internal argument a nonfinite clause with an implicit subject (PRO)[13]: with this respect, a subtle metric of complexity is given by the lexical properties of control[14].

Under the «control theory» developed in the Government and Binding framework (Chomsky, 1981), verbs can be distinguished into two main categories according to the constraints they place on the interpretation of the embedded PRO, which can be coindexed with the matrix subject or the matrix object: the two options are exemplified, respectively, by *promise* (ex. 1) and *order* (ex. 2) type verbs. There is also a third class of verbs, which the verb *ask* is a prototypical case of, admitting both the syntactic options, namely the object-related interpretation of PRO, when *ask* bears the meaning of a 'request', as in (3a), and the subject-related interpretation of PRO, when *ask* is used to express a question, as in (3b).

(1) The editor$_i$ *promised* the journalist$_j$ PRO$_{i/*j}$ to write the article.
(2) The editor$_i$ *ordered* the journalist$_j$ PRO$_{*i/j}$ to write the article.
(3) a. The editor$_i$ *asked* the journalist$_j$ PRO$_{*i/j}$ to write the article.

[13] In generative grammar models, PRO is the formalism to indicate the null subject of infinite clause and it has to be intended as an empty category bearing the features [+anaphoric, + pronominal].
[14] Verb control information and its role in establishing verb complexity hierarchies has been outlined, among others, by Laura Ciccarelli e Marica De Vincenzi (1996). Their study has been particularly inspiring for the discussion contained in this paragraph, as the authors, moving in a psycholinguistic perspective, revised some linguistic complexity factors with the precise intent of showing the weakness of traditional readability formulae (cf. §2.2).

b. The editor$_i$ *asked* the journalist$_j$ what PRO$_{i/*j}$ to do.

In this case, much evidence in support of the existence of a diverse degree of complexity behind these structures stems from language acquisition research. In particular, since a seminal work of Carol Chomsky (1969) with English-speaking children from 5 to 10, it was noticed that object control sentences, such as (2), were mastered earlier than subject control ones (1); the latter, in turn, preceded the acquisition of verbs admitting both the control structures (3), whose stable comprehension was not achieved before 10. This pattern of development has been replicated cross-linguistically (for a review, cf. Guasti, 2002: 347-371) and, interestingly, some studies have revealed that it has a similar counterpart in adult online comprehension (Frazier *et al*., 1983)[15].

According to Chomsky's interpretation, the observed delay with subject control structures might be expected under a notion of linguistic complexity formalized as follows:

> «The syntactic structure associated with a particular word is at variance with a general pattern in the language»[16].

Within the specific domain of control, the «general pattern» of English is the object-referred interpretation of the null subject of the infinitive clause. Such an interpretation is constrained by the Rosenbaum (1967)'s Minimal Distance Principle (MDP), which, in its turn, represents one of the earliest attempts to define the concept of syntactic "locality" (cf. §1.3.1). In fact, not only are object control structures more frequent in English (the "order" type class is bigger than the "promise" type one), but their computation is also favoured for economy reasons: indeed, as the matrix object is linearly closer to the complement PRO, it is likely to be higher activated in short memory, and thus preferred as controller over the more distant matrix subject.

Chomsky claims that children, before achieving the adult-like competence in their native language, will apply MDP both in the compatible (2) and incompatible settings (1), presumably because this is a strategy minimizing memory requests[17].

[15] But, for different findings, see Boland et al. (1990).
[16] C. Chomsky (1969: 6-7).
[17] A different proposal to account for the children's difficulty with 'promise' type verbs has been recently advanced by Belletti and Rizzi (2012); here the authors argue that control, similarly to passive, is another domain amenable to give rise to an 'intervention' effect ruled out by Relativized Minimality (cf. 1.3.4); in

After learning that "promise" type verbs represent an exception to MDP, a further maturational stage is needed to master the "ask" type verbs. The latter case, indeed, faces them with an "incoherent" exception, as it allows both the object and the subject control reading; as we noted before, processing a verb that admits more potential argument structures increases the perceived difficulty of the sentence in adults, too.

B) **Argument mapping effects**

An additional factor adding to verb processing complexity is related to the *argument mapping*, namely the structural realization of the thematic roles (e.g. agent, patient, goal) as syntactic arguments (e.g. subject, object, indirect object).

Such a correspondence is regulated by the *thematic hierarchy*, which is a fundamental theoretical formalism conceived to establish the ranking of the thematic roles associated with verbal entries. A corollary assumption is that the most prominent the role in the hierarchy, the higher the syntactic position it will be mapped onto[18].

Despite different hierarchies have been proposed in the literature according to diverse reference frameworks[19], a common tenet is that, whenever present, the Agent role in unmarked contexts will be assigned to the higher syntactic position[20]. When the Agent is missing, a different role can be promoted to the subject position: this is what occurs, for instance, with psychological verbs (psych-verbs).

Psych-verbs are verbs denoting a mental state, which require two semantic arguments: an *Experiencer*, i.e. the argument experiencing the psychological state denoted by the verb, and a *Theme*, i.e. the object/content of the experienced state. On the basis of the argument they select as sentential subject, they can be distinguished into two main categories: subject-Experiencer verbs, like *fear/admire/dislike* (ex. 4) and object-Experiencer verb, such as *frighten/amuse/distress* (ex. 5).

order to avoid intervention, additional movement operations within the argument structure are required, which are costly and thus might become available only at a later developmental stage.

[18] A stricter version of the mapping issue is known as the *Uniformity of Theta Assignment Hypothesis* (UTAH), which assumes that «identical thematic relationships between items are represented by identical structural relationships between those items at the level of D-structure» (Baker, 1988: 46).

[19] Cf. Levin (2006) for a comprehensive review.

[20] The "external argument [*Spec, VP*]" position, the "subject" position or the "nominative case" position, depending on the referential syntactic theory.

(4) My students *feared* oral exams.

(5) Oral exams *frightened* my students.

Following Belletti and Rizzi (1988)'s proposal in the generative grammar framework, in *fear*-type verbs the theta-role mapping is still transparent: the Experiencer ('my student'), like the Agent, is assumed to be based generated in the external argument position, resulting in a typical transitive structure. On the contrary, *frighten*-type verbs exhibit a thematic mapping mismatch, which is considered to derive from syntactic movement: more specifically, the Theme has undergone a displacement from the position where it is originally projected (the internal complement of vP) to the external argument position (the specifier of vP)[21]. This is an instance of A-movement, i.e. a movement where a syntactic element (in this case the nominal object phrase (NP)) lands to a thematic position, as in the case of passive (see §1.3.1); nonetheless, differently from passive, here the movement is triggered by the lexical semantic properties of the verb rather than by morpho-syntactic issues.

For the purpose of our discussion, it is worth underlining that the more complex representation underlying object-Experiencer verbs[22] has a counterpart in language processing data, both in normal and impaired speakers.

For what concerns the former, Brennan and Pylkkänen (2010) conducted a study on the comprehension of psych-verbs in adult English speakers, by using both behavioral and neuroimaging measures[23]. A significant delay was reported in reading sentences with an object-Experiencer verb with respect to sentences containing a subject-Experiencer verb[24], although no difference at the level of neural activity was found between the two conditions.

With respect to neurolinguistics data, the higher processing cost associated with object-Experiencer verbs was assessed in a study of Manouilidou *et al.* (2009) testing English speakers affected by Alzheimer's disease (AD). In the proposed task, the participants were

[21] In 2012, the authors slightly changed their original proposal by assuming that the movement does not involve the only NP but instead the whole verbal chunk containing the internal object, thus an instance of a "smuggling" operation.

[22] Note that a difference in terms of complexity between the two categories of psych-verbs is also expected under different theoretical approaches, such as decompositional semantics analyses. For instance, Peseztsky (1995) claims that object-Experiencer verbs are more complex because their underlying representation has a richer feature detail: in particular, he postulates the presence of an additional causative semantics, realized by a Causer marked with zero morphology which fills the subject position.

[23] The online comprehension of psych-verbs was assessed through the self-paced reading technique for what concerns the behavioral measures, and by using magnetoencephalography for evaluating the neuronal correlates.

[24] The difference in reading times was registered on the word immediately following the target verb.

presented with a sentence frame containing a missing verb and asked to read, and then to fill in, the given sentence by choosing the right verb within a list of four potential alternatives.
The experimental stimuli were organized into six conditions, resulting from the combination of two variables: the verb type (i.e., subject-Experiencer/object-Experiencer/subject-Agent verb) and the verbal voice (active vs. passive)[25]. The latter alternative was introduced with the precise intent of controlling the role of syntactic movement – a well-established source of processing difficulty in brain damaged populations, such as Broca's aphasics (§1.3.3) – so that to make it possible to verify the existence of a "pure" effect of canonicity determined by the thematic hierarchy rather than by syntactic manipulations.

This is exactly what has been found. Specifically, while AD patients had no problems with passive compared to active sentences (thus showing a different pattern than Broca's aphasic speakers), they exhibited a selected deficit with active sentences displaying a non-canonical argument realization. Interestingly, not only their performance was lower when asked to complete object-Experiencer verb frames (ex. The thunder *frightened* the children) rather than their subject-Experiencer counterpart (ex. The children *feared* the thunder), but they were also slightly impaired in the latter condition compared to the Agent active items. According to the authors, also the argument realization in subject-Experiencer verbs is atypical, in that it preserves the thematic hierarchy, but still deviates from the "default" realization assigning the Agent role to the first NP; thus it is amenable to be more difficult to process for brain damaged readers.

1.3 Syntactic complexity

We now turn the attention to the process of sentence comprehension, whereby the human language processor (parser) assigns the grammatical role to each incoming word in order to construct a propositional representation that constitutes the meaning of the sentence.

At this level, the problem of what makes a sentence more or less difficult to comprehend can be tackled by considering both the properties of syntactic computations and how these properties affect sentence parsing.

[25] The sentence frames were distributed into the following conditions: 1) subject-Experiencer/ active verbs (e.g. *fear*); 2) object-Experiencer/active verbs (e.g. *frighten*); 3) subject-Agent/ active verbs (e.g. *save*); 4) subject-Experiencer/ passive verbs (e.g. *was feared*); 5) object-Experiencer/passive verbs (e.g. *was frightened*); 6) subject-Agent/ passive verbs (e.g. *was saved*).

The comprehension of a sentence is indeed a real-time process, which is primarily affected by human memory limitations. More specifically, the parser has been shown to rely on subsidiary mechanisms (e.g. the working memory span[26]), which play an influence since the very beginning of sentence parsing (the so-called "first-pass parsing"). These mechanisms lead the parser to assemble incoming lexical material into the "simplest" structure, i.e. the structure minimizing the computational resources required to generate a meaningful representation of the sentence. According to this view, the human processor is said to be driven by "economy" principles.

But what makes a syntactic computation easy and, conversely, difficult to carry out in light of such a general architecture and extra-linguistic constraints? An answer to this question, albeit not comprehensive, requires us to introduce a set of formalisms that current syntactic theory has devised to characterize the properties of grammatical structures.

1.3.1 Some theoretical background in a *minimalist* framework

Since the earliest generative models, it has been postulated that syntactic computations are diverse in nature, as they occur to satisfy a different set of properties displayed by lexical items. Basically, two types of computations have been distinguished, which have come to be qualified under the current minimalist labels of "Merge" and "Move" (Chomsky, 1995)[27].

Merge is the "core" combinatorial operation, which is responsible of creating the basic syntactic structure (similar to the structure that in traditional generative models was referred to as the "deep structure") by picking two elements from lexicon and joining them together, in order to satisfy the selectional requirements of the head, i.e. the element which projects its category (e.g. a verb): the outcome of Merge is a higher-order category, which in turn should be merged recursively with another lexical category, yielding progressively the syntactic tree of a whole sentence. It is the Merge operation that allows the thematic role assignment to be established (cf. paragraph 1.2.1), by combining the lexical head (e.g. a two-place verbal predicate) with its arguments (e.g. a NP object and a NP subject), as in (6).

[26] Cf. the influential paper of G.A. Miller (1956) on "the magic number seven", in which the author fixed the working memory capacity for processing information to seven (plus o minus two) units. On the role of working memory in language comprehension, see, among others, Daneman and Carpenter (1980);

[27] A further development of the Minimalist Programme (Chomsky, 2001) reduces the two different computations to the only "merge" operation, by distinguishing it into "external" and "internal" according to the source (the lexicon or the structure) from which the element to be merged is selected.

(6) Rachel wrote a paper.

Move, instead, applies to the output of Merge and is responsible of modifying the structure built thus far, by selecting previously inserted lexical items and moving them to higher positions of the tree, thus giving rise to non-local syntactic dependencies; as a result, the empty position is filled by a trace[28] (i.e. the symbol <*t*> in the following examples), which has to be coindexed with the landing site position of the moved element by means of a chain, so that the sentence can be properly understood.

Like Merge, also Move is "feature-driven", namely triggered by specific types of lexical properties that must be satisfied within dedicated syntactic positions in order to fully interpret the meaning of each element in the structure. More specifically, two types of features can trigger the displacement of a lexical item: morpho-syntactic features (e.g. verbal voice, case) and scope-discourse features (e.g. topic, focus, etc.). The different nature of these features is responsible of what, in standard X-bar theory (Chomsky, 1957, 1981) were defined, respectively, as instances of A and A' movement, according to the landing site position of the moved item[29]: A movement targets an argument position, namely a position in which theta-roles can be assigned, and it allows satisfying the properties of an item (a whole XP, e.g. a noun and its complements) at the morphology-syntax interface; A' movement, instead, is a movement to a non-argument position, namely a left-periphery position (Rizzi, 2004), where the displaced element (e.g. a specifier bearing [+wh] features) can receive a scope-related interpretation.

Examples of each case are provided below. In sentence (7), the transformation of an active into a passive sentence is a typical case of A movement affecting the internal argument of the verb ("the paper"), which is moved from its canonical position (the complement of V) to the external argument position (the subject position).

(7) The paper was written by Rachel.

[28] The more recent "copy-theory" of movement (Chomsky, 1995) defines traces as complete, but unpronounced, copies of the moved element. In the examples presented further we will adopt the traditional model of traces.

[29] An additional type of movement affects lexical heads; a typical example is the "Subject-Auxiliar inversion" in English, in which the auxiliary verb raises over the subject to form interrogative questions. Here we focus on A and A' movement because of their implications in defining different typologies of complex sentences.

Conversely, (8) exemplifies a particular type of non-local dependency of A'-type (also called a "filler-gap" dependency in psycholinguistics literature), in which the *wh*-operator ("what") has undergone movement from a lower position of the tree (i.e. the internal argument position of the embedded verb) to a dedicated position in the left-periphery.

In order to establish the required chain, the parser has to keep in memory the filler until the gap position is processed: as we will see (§ 1.3.3), such an operation may prove to be particularly costly according to the hierarchical structure and the featural properties of the elements involved.

(8) I wonder *what*$_i$ Rachel wrote $_{<t_i>}$?

Although displacement operations represent the norm rather than the exception in natural language, a distinguishable property of syntactic computations is to be subject to locality conditions, which impose restrictions to the typology of elements entering into a syntactic chain. Locality is a powerful explanatory device capturing fundamental aspects of linguistic structures not only limited to the syntactic domain (see Rizzi, 2013 for a review). Under its specific formulation known as Relativized Minimality (RM), originally proposed in Rizzi (1990), locality can be regarded as an economy principle, since it allows the parser to limit the portion of the tree within which a given local relation has to be computed. According to RM, indeed, a relation between two elements, X and Y, is banned when an element Z, matching the same features of X, intervenes between X and Y, namely in a configuration like (9).

(9) RM: …X…Z…Y…

It is worth underlying here that RM is not a concept with a pure formalistic appeal; as demonstrated by much current research both in children and adult sentence processing, it also bears relevant implications in modulating syntactic complexity effects. However, to fully appreciate how this principle can contribute to establish a "hierarchy" of difficult sentences, we need to introduce some more syntactic background.

Simplifying somewhat, we can say that to "count" as a potential intervener in a syntactic dependency, Z has to be fully specified by the same set of features licensing X. These features are classifiable into four classes, according to the following typology[30]:

(10)

a. Argumental: person, number, gender, case

b. Quantificational: Wh, Neg, measure, focus..

c. Modifier: evaluative, epistemic, Neg, frequentative, celerative, measure, manner..

d. Topic

In sentence (8), we could see the establishment of a chain between the target element ("what") - which we can now qualify as a specifier licenced by Quantificational features - and its trace in the object position. Such a sentence is wholly acceptable, since no potential intervener for RM occurred in the relevant chain: indeed, the subject of the embedded clause ("Rachel") is an element of a different nature, namely a specifier licenced by Argumental features.

The same does not hold in (11), where the internal subject (i.e. "who") belongs to the Quantificational class itself and, as predicted by RM, blocks the required relation between the *wh*-object and its trace.

(11) * I wonder what$_i$ *who* wrote $_{<ti>}$

For the purpose of understanding syntactic complexity from a user-based perspective, we now review some data providing empirical evidence that both Merge and Move operations, when occur within particular syntactic configurations, can affect the ease with which the surface sentence is interpreted. We will finally draw the conclusion (cf. § 1.3.4) that greater processing effort is required by those sentences whose underlying derivation somehow resembles (9), i.e. the configuration prohibited by the RM principle.

[30] The features covered by this typology define in turn precise "slots" within the syntactic tree of a sentence. This is the core assumption of the *Cartographic Approach*, a research trend devoted to providing fine-grained maps of syntactic representations. For in-depth explanations, the reader is referred to Belletti, ed. 2002, Cinque, ed. 2002, Rizzi 1997, ed. 2002, 2004 and related work.

1.3.2 Ambiguity and multiple embeddings

One of the most investigated sources of syntactic complexity is represented by ambiguous sentences, namely sentences admitting two (or more) readings according to the underlying structure.

The resolution of syntactic ambiguity, which is a pervasive phenomenon in language, has enjoyed a great interest in language-related disciplines, especially psycholinguistics and computational linguistics[31], as it opens up a window into the mechanisms underlying online human language processing and their possible simulation. With this respect, psycholinguistic research has identified three major "resource-saving" strategies that the parser adopts to resolve ambiguity: Minimal Attachment, Late Closure and Minimal Chain Principle.

Minimal Attachment principle (Frazier and Fodor, 1978) accounts for the tendency of the parser to build the structure requiring the fewest number of syntactic nodes, and which is consistent with the input. In a sentence like (12), Minimal Attachment predicts a stronger preference for interpreting the prepositional phrase ("with the bag") as an adjunct of the verb – i.e. a position already available in the tree after the verb is processed – rather than as a modifier of the NP object.

(12) The little boy covers the doll with the bag.

According to the Late Closure strategy (Frazier and Fodor, 1978; Frazier, 1979), similarly captured by the Right Association Principle (Kimball, 1973) and the Recency principle proposed by Gibson *et al*. (1996), when two potential interpretations can be derived through the same number of syntactic nodes, the parser prefers to merge incoming lexical items to the most recently built phrase or clause.

For instance, in a sentence like (13), both a high and a low attachment of the relative clause ("that arrived yesterday") are grammatically possible: yet, Late Closure makes the listeners/ readers to prefer the latter, i.e. the interpretation of the relative clause as modifying the second NP ("her friend").

[31] Cf. De Vincenzi and Ciccarelli (2004: 95) [*translation mine*]: «The study of ambiguity within these fields is so overspread that whoever approaches psycholinguistics and computational linguistics for the first time would have the wrong impression that the aim of psycholinguists and computational linguists is limited to understand how the human language processor works out ambiguity».

(13) Mary bought a present to the nephew of her friend that arrived yesterday.

Finally, Minimal Chain Principle ("MCP"), (De Vincenzi, 1991), plays a role in the resolution of ambiguities derived from non-local syntactic dependencies of A' type. As we mentioned in the previous paragraph, such constructions require the parser to establish a chain between the filler and its trace in the gap position, so that the former can receive the correct semantic interpretation in the discourse.

Let us consider the case of *wh*-questions in Italian, a language that allows the subject to occur in a post-verbal position; in light of this property, a sentence like (14) is ambiguous since the *wh*-element ("Chi") can be interpreted both as the subject and the object of the clause[32], according to the syntactic chain it is assumed below.

(14) Chi ha chiamato Rachele?
'Who has called Rachel?'

(14a) Chi_i *pro* ha chiamato $_{<ti>}$ Rachele? → wh-object extraction

(14b) Chi_i $_{<ti>}$ ha chiamato Rachele? → wh-subject extraction

Based on large experimental evidence showing increased reading times for *wh*-object extractions, MCP argues that the parser will prefer to interpret the filler in the first empty position that is structurally available[33], i.e. the subject position in the examples above, so that to reduce the time the filler has to be stored in memory.

But in what way these strategies can ultimately affect sentence complexity? To clarify this point, we need to consider that language presents listeners/readers with both full and temporary ambiguity.

[32] The ambiguity depends on the fact that both the first and the second NP have compatible number and person features with the verb, which in Italian agrees with the subject with respect to these features.
[33] Minimal Chain Principle: «Avoid postulating unnecessary chain members at S-structure, but do not delay required chain members.» (De Vincenzi, 1991).

In the latter, the so-called "garden-path" sentences[34], the ambiguity is resolved before the end, after a certain word is processed. To make an example, let us suppose that (12) would continue as (15): in this case, the feminine pronoun within the last PP ("on her hand") rules out the structure created so far, following Minimal Attachment.

(15) The little boy covers the doll with the bag on her hand.

Nevertheless, the first-pass parsing is highly deterministic. This means that, when faced with temporary ambiguity, the parser does not delay the analysis until the "disambiguating" cue is encountered, nor it keeps in memory all the possible representations that are temporarily compatible. Instead, it goes on with the analysis that is preferred by the heuristic structural principles, in order to process incoming material more quickly. As a consequence, the later the ungrammaticality is detected, the more difficult the reanalysis process and challenging the comprehension of the whole sentence.

However, not only ambiguous sentences cause extreme processing difficulties. There exists a variety of structures whose interpretation lays on a bigger storage of computational resources.

Let us consider first the (external) merge operations, which build larger phrases from smaller ones. At this level, a well-investigated index of syntactic complexity is represented by the "amount of structure" that each incoming word carries with itself, calculated as the number of syntactic nodes that the parser has to project in order to integrate each new word into the given structure. Several metrics have attempted to operationalize sentence complexity by means of a "node-counting" algorithm: it is the case of Yngve (1960), who calculates the parse-tree depth as the number of incomplete syntactic dependencies on the stack; Miller and Chomsky (1963), who introduced the "node to terminal node ratio"; or Frazier (1985), focusing on local nonterminal count (cf. Szmrecsányi, 2004 for a review).

To exemplify what these metrics account for, we can consider, for instance, the growing complication in reading the following sentences, which is directly related to the number of recursive nodes within the sentential subject phrase.

[34] The literature on garden-path sentences is extremely rich. Some of the most influential works are Bever (1970); Frazier and Fodor (1978); Stowe (1989). See Gibson (1998) for a review.

(16) That those parents always complain to the headmaster is annoying.

(17) That the fact that those parents always complain to the headmaster is taken for granted is annoying.

(18) That the fact that if the teacher assigned more homework those parents would complain to the headmaster is taken for granted is annoying.

Note that the overload effect here is different from the garden-path one, as no ambiguity is encountered; yet, it is explainable on a similar memory-based account, so that the higher is the number of incomplete syntactic dependencies that the parser has to keep track of, the harder the sentence will be perceived by the listener/speaker.

It can be argued that what makes (18) more difficult to process is the semantic density of the sentence, calculated as the number of propositions involved[35]. This is surely an additional factor of complexity (cf. §1.4); however, if we compare (18) to (19), the latter is intuitively less demanding, thus proving that, despite the same length (in terms of number of words)[36] and propositions, the processing of self-embedded rather than right-branching nodes is an influential variable of syntactic complexity to be considered.

(19) It is annoying the fact that it is taken for granted that if the teacher assigned more homework those parents would complain to the headmaster.

Next paragraph will consider more in detail the role of syntactic movement as a metric of complexity by focusing on the effects triggered by diverse types of incomplete syntactic dependencies on sentence elaboration.

1.3.3 Moved-derived sentences

As we discussed above, if incomplete syntactic dependencies have important implications for human sentence processing, it is also true that not all unbounded dependencies are costly to the same extent. With this respect, a metric to evaluate syntactic complexity is offered by

[35]A proposition can be viewed as an idea unit, i.e. a semantic concept in which the meaning of the sentence it is not conveyed by the exact wording and syntax but it is translated into a relationship between a predicate and at least one argument (Van Djik and Kintsch, 1973).

[36]Actually, the "easier" version in (19) contains even more words than its counterpart in (18) (i.e. the two obligatory expletive pronouns at the beginning of the main and the first embedded clauses).

movement, which in formal models is responsible of displaying elements in a position of the tree that is different from the position in which they were originally merged (§ 1.3.1).

Although the possibility of establishing syntactic relations between non-adjacent elements is a distinctive property of natural language, its impact in comprehension has been acknowledged since the early theories of sentence processing. It is worth mentioning here the *Derivational theory of complexity* (DTC), (Miller and Chomsky, 1963; Miller and McKean, 1964), based on early generative transformational models, which gave rise to a wide experimental program pursuing the hypothesis that sentence complexity was a linear function of the number of derivational steps needed to convert the deep structure into the surface structure. In particular, while some of these transformations were considered as obligatory, since in their absence the final sentence would have been ungrammatical (e.g. the derivational steps required to produce the subject-verb agreement), other ones were deemed optional (e.g., the passivization rule transforming an active sentence into a passive one). According to the DTC, the more a sentence displayed optional transformations, resulting in a deviation from the simple structure (i.e. kernel), the more complex was to process by the reader/listener.

However, large experimental evidence rejected the psychological plausibility of the DTC (Fodor, Bever and Garrett, 1974 for early accounts) by highlighting the "imperfect" correlation between the postulated amount of transformations and the resulting processing effort.

Just to make an example, a well-studied processing phenomenon is the asymmetry between subject-extracted (SRC) (ex. 20) and object-extracted relative clauses (ORC) (ex. 21), where the former are easier to process by a number of measures taken from adult performance, such as phoneme monitoring, online lexical decision, reading times, response accuracy[37], as well as with respect to cross-linguistic data from language acquisition research[38].

(20) SRC: *The journalist$_i$ who* <t_i> *attacked the senator* met the editor.
(21) ORC: *The journalist$_i$ who the senator attacked* <t_i> met the editor.

[37] This is a well-documented psycholinguistic phenomenon; see, among many others, King and Just, 1991; Warner and Maratsos, 1978; Just *et al*., 1996; Traxler *et al*., 2002, Gordon et al., 2004.
[38] See, among many others, Belletti and Contemori (2010), Friedmann, Belletti and Rizzi (2009); Arosio *et al*., (2005) and references therein.

Yet, the two sentences go through the same derivational process (i.e. a *wh*-extraction of the NP phrase out of the embedded sentence), and what changes is only the structural position (i.e. external subject or internal object) from which the movement takes place within the embedded clause[39]. To understand this phenomenon, as well as some even more subtle asymmetries within the only domain of object *wh*-extractions (see further), we thus need to consider the interplay between movement and other properties of syntactic structures.

As a consequence of syntactic movement, the sentence may undergo a non-canonical thematic assignment (§1.2.1), with the patient/theme preceding (both linearly and hierarchically) the agent phrase, as it occurs e.g. in passives (22), object *wh*-question (23) or the above mentioned object relative clauses, both in right-branching (24) and self-embedded position (25).

(22) The journalist was attacked by the senator
(23) Which journalist did the senator attack?
(24) The editor met the journalist who the senator attacked.
(25) The journalist who the senator attacked met the editor.

The correct interpretation of these sentences, and specifically of the reversible type[40], necessarily relies on pure grammatical devices; in particular, what is at play is the chain interpretation mechanism, which allows the moved constituent to be coindexed with its trace in the first-merged position (§1.3.1).

While for moved-derived sentences preserving the standard thematic assignment (e.g. subject relative clauses) the right interpretation might still be reached via adopting economic heuristics, such as that of attributing the role of agent to the first noun processed (Slobin, 1966), the same strategy would not be felicitous to deal with their non-canonical counterparts, making the latter more problematic to process. This effect has been widely confirmed by neurolinguistics research, particularly with aphasic speakers, where the misinterpretation of (reversible) non-canonical sentences is included among the syntactic

[39] In the examples above reported, the relative clause is marked in italics. The gap position within the relative clause is indicated by the trace, which is coindexed with the head of the relative head. It should be noted that this is an oversimplified representation, which is compatible with both a raising and a matching analysis of relative clauses. The interested reader can see (Bianchi, 2002) for in-depth explanations.

[40] In a reversible sentence (e.g. The boy is hugging the girl), the arguments bearing the role of subject and object may be interchanged without resulting in a semantically odd sentence. It is only the syntactic structure indeed that allows the correct thematic role assignment, differently from what occurs with non reversible sentences (e.g. The secretary is writing the document), which can rely on semantic cues.

"signatures" of Broca's agrammatism[41], but it also been assessed in online normal processing, and interestingly with respect to irreversible non-canonical sentences, too (Ferreira, 2003).

However, non-canonical sentences represent a "macro" category of difficult sentences, which is internally further distinguishable according to the type of underlying movement, elements and portions of the tree involved. It is worth considering such variables, as they have proven to be involved in modulating speakers/readers' processing efforts, thus allowing us to introduce a more detailed hierarchy within the category of "complex sentences".

One discriminating dimension is represented by the type of movement, i.e. movement targeting a thematic or a non-thematic position (§1.3.1). In this regard, sentence (22) is different from the group in (23)-(25) because only the former entails the movement of the internal argument (i.e. the direct object of the verb) to a thematic position (i.e. the subject position). Such a kind of (A type) movement may be considered simpler than movement to the left periphery of the clause: not only it affects a portion of the tree which is hierarchically closer to the source position of the moved element, but it also prevents the sentence from undergoing the "intervention" effect which, as we will see in a short while, is responsible of the more problematic computation affecting object A'-dependencies.

Again, the validity of this assumption seems to be tenable on empirical grounds: agrammatic speakers e.g. tend to preserve the access to lower nodes of the structure with respect to higher nodes, which are more susceptible to impairment (Friedmann and Grozinsky, 1997). Also in typically developing children, although passive is mastered later than active form, it still precedes the stable acquisition of other kinds of complex non-canonical sentences, such as object relative clauses. Analogously, despite in a different modality (i.e. production), recent findings from elicited production with Italian adult speakers (Belletti and Contemori, 2010) revealed that the use of passive, within given contexts prompting the relativization of the direct object, is largely preferred over the production of standard object relative clauses.

However, also within the group of sentences in (23)-(25), which all contain an object A'-dependency, we can come up with further distinctions in terms of processing complexity. Let us remember that the interpretation of non-local dependencies of the filler-gap type requires the parser to establish a chain between the surface position of the moved constituent (target) and its original thematic position (gap).

[41] Caramazza and Zurif, 1976; Bastiaanse *et al*., 2002; Grodzinsky, 1990; Friedmann and Shapiro, 2001; Thomson and Shapiro, 2007; Grillo, 2008, among many others.

While in subject A'-dependencies (e.g. the subject relative clause in (20)), the two positions are co indexed by means of a shorter chain, which favors the application of economic processing routines such as Minimal Chain Principles (§1.3.2), the same does not hold for object A'-dependencies, where the greater distance between the two non-local elements poses extra load on memory.

In this regard, a significant attempt to formulate a syntactic complexity function rooted on psycholiguistic evidence has been offered by the *Dependency Locality Theory* by Gibson (1998, 2000), which gauges sentence processing costs by considering two components: a "storage" component and a "structural integration" component. The former establishes what quantity of memory units is needed to keep track of a partial input sentence and it corresponds to the minimum number of words that is required by each syntactic head to complete the sentence in a grammatical way; the latter enables the incorporation of new words into the previous structure and it depends on the number of elements intervening between the target and the gap and introducing new discourse referents (roughly, nouns and verbs but we will come back to this issue later). Under this model, the observed asymmetry between subject vs. object A' dependencies seem to derive from the fact that the latter undergo higher processing costs for both these components.

This is exemplified by the two profiles reported below. For what concerns the 'storage cost' (cf. 26), when the point of maximum processing load is reached (i.e. the relative pronoun in the SRC and the second occurrence of the determiner in the ORC), the ORC will exceed of 1 memory unit the SRC. This is because, at this point, the parsing of the ORC meets four incomplete syntactic dependencies: 1) the NP ("the journalist") needs a verb; 2) the relative pronoun (which signals the presence of a gap in the structure) informs the parser to expect (at least) two more heads, i.e. another verb and an empty category to host the trace and 3) the determiner after the pronoun needs itself a noun.

(26) **Storage cost** for SRC vs ORC (Gibson 1998, 2000)

(20) SRC: The journalist who attacked the senator met the editor.

2 1 3 2 2 1 1 1 0

(21) ORC: The journalist who the senator attacked met the editor.

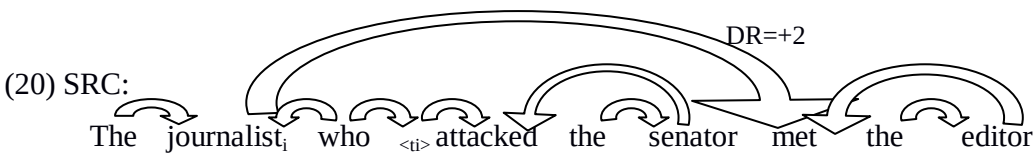
2 1 3 4 3 1 1 1 0

Also with respect to the 'integration' component, the ORC is expected to be more demanding than its subject counterpart. What is crucial here is the "spatial" dimension of syntactic complexity, i.e. the number of computational states to be crossed in the course of the derivation[42]. As we slightly mentioned before, in Gibson's (1998, 2000) model this property is translated into a *distance-based* measure, according to which the cost of performing a structural integration (SI) between a head and its dependent is proportional to the number of new discourse referents (DR) in the intervening region. The "likelihood" of an element to count as an intervener in the relevant chain (and thus to increment integration costs) is claimed, in turn, to be proportional to the degree of accessibility that the element bears in the discourse: more specifically, both the head verb of a VP and the head noun of a full NP are considered as the more demanding to process, since they introduce new referents in the discourse; proper names are "less heavy" than full NPs, since they refer to non-focused entities, while pronouns are the least heavy ones, as they refer to focused entities.

If we compare the two profiles in (27), it can be noted that the SRC undergoes the maximum integration cost at the main verb ("met"): at this level indeed one memory unit is required to construct a new discourse referent (which corresponds to the verb itself), and two additional ones (corresponding to the intervening referents, i.e. "attacked" and "senator") are also necessary for the integration of the subject with the main verb.

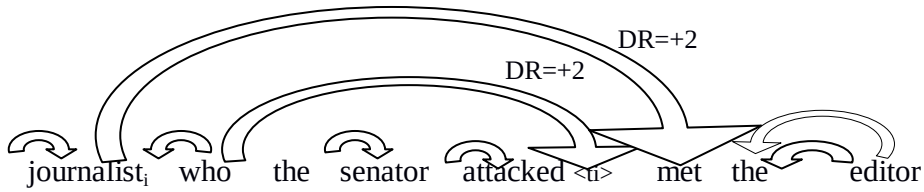
In the case of the ORC, these costs are twofold since two non-local dependencies have to be worked out (i.e. one between the matrix subject and the matrix verb and one between the head of the relative clause and its object empty category), which both consume the same quantity of computational resources, given the same number of intervening discourse referents.

(27) **Integration cost** for SRC vs ORC (Gibson 1998, 2000)



[42] See also Chesi (2012:68-69) and Chesi and Moro (2013) for a more in-depth explanation.

(21) ORC:



	The	journalist$_i$	who	the	senator	attacked <t_i>	met	the	editor
DR	0	1	0	0	1	1	1	0	1
IC	0	0	0	0	1	3	4	0	1

Let's now go back to the group of sentences in (23)-(25), which are repeated below.

(23) Which journalist did the senator attack?
(24) The editor met the journalist who the senator attacked.
(25) The journalist who the senator attacked met the editor.

As we can observe, although they all display an object A'-dependency that is more challenging to elaborate for the human parser given the reasons so far discussed, only in (24) and (25) the movement takes place from an embedded position. Thus, on the basis of the parse tree depth and the number of propositions involved (cf. §1.3.2), the latter are perceived as more difficult than the object *wh*-question in (23).

Once again, the inspection of neurolinguistics data allows us to endow such grammar-based predictions with empirical support. An interesting study was conducted by Thompson and Shapiro (2007) with Broca's agrammatic speakers, where it was shown that training more complex sentences improved the comprehension of untrained less complex ones (but not the opposite pattern), provided that the same type of syntactic movement was implicated by the trained/untrained material. In accordance to this hypothesis, the treatment of object relative clauses successfully generalized to untreated object *wh*-questions, but not vice versa, thus proving the higher complexity of the former over the latter.

Finally, despite the same number of embeddings, the sentence containing a self-embedded object relative (25) is even more complex than its counterpart in (24), where the object relative clause is right-branching.

Since the former, but not the latter, requires the parser to establish two non-local dependencies (i.e. the first one between the matrix subject and the matrix verb and the second one between the relative head and the object trace in the embedded relative clause), sentences like (25) will face higher "storage" and "integration" costs, making them more prone to break down.

Just to recap the main issues so far discussed, we can say that, from the point of view of human sentence processing: i) moved-derived sentences, which are amenable to display canonicity effects, are responsible of higher processing difficulty; ii) when the sentence contains a non-local dependency featuring object *wh*-extraction, both the whole parse tree depth and the embedded position with respect to the matrix clause have to be considered as possible variables implied in sentence complexity.

Before concluding this section, which has been devoted to re-examining grammar-based accounts of complexity from the psycholinguistic perspective, in the next paragraph we will take into account an intriguing pattern regarding the comprehension of different typologies of object non-local dependencies, which seem to suggest the need of introducing more fine-grained metrics to gauge sentence complexity.

1.3.4 A "featural approach" to syntactic locality as a metric of sentence complexity

As we discussed in the previous paragraph, the asymmetry between subject vs. object non-local dependencies is a well attested processing phenomenon, which can be explained by a *distance-based* metric capable of predicting the higher costs posed by unbounded material across intervening elements.

However, all other things being equal (i.e. hierarchical depth and embedding position), some findings from adult sentence processing (Gordon *et al*., 2001, 2004), as well as offline children comprehension (Friedmann *et al*., 2009) reveal that the processing efforts[43] posed by object A'-dependencies are highly sensitive to the "featural make-up" of the elements involved in the relevant chain, i.e. the moved constituent and the intervening element.
With this respect, let's consider the following sentences:

[43] Processing costs are translated into higher reading times for adults and change-level performance for children.

(28) The journalist$_i$ who the senator attacked $_{<t_i>}$met the editor.

(29) It was Mark$_i$ who John attacked $_{<t_i>}$

(30) The journalist$_i$ who Rachel attacked $_{<t_i>}$ met the editor.

(31) The journalist$_i$ who I attacked $_{<t_i>}$ met the editor.

(32) The one$_i$ who the senator attacked $_{<t_i>}$ met the editor.

What emerges from the above mentioned studies is the existence of a "hierarchy" of processing complexity in the domain of object non-local dependencies, where the "peak" is represented by sentences like those in (28) or (29), i.e. sentences with a relative clause[44] in which both the target and the intervener are realized by the same nominal expression (here, respectively a full NPs and a proper noun). On the contrary, comprehension ameliorates when these sentences undergo a slightly modification affecting either the target or the intervener, so that the two elements turn out to be distinguished with respect to their lexical realization. This occurs, e.g. in (30) and (31), where the embedded subject is realized, respectively, as a proper name and a pronoun, but also in the case of a free relative like (32), where the intervener is a full NPs and the relative head is a demonstrative pronoun.

Interestingly, a complexity metric like the one proposed by Gibson, which accounts for the higher integration costs in terms of the discourse referentiality of the intervener (§ 1.3.3), does not appear to be fully adequate to explain the whole set of empirical data. In particular, it "fails" in (29) and (32): the former is predicted to be easier since the crossing element is a proper name, which is deemed more accessible than full NPs, but this is not confirmed by experimental evidence; the latter instead is assumed to be as difficult as (28), given the presence of a "heavy" full lexical noun in the relevant intervening region, but again this prediction is rule out by performance data.

A promising alternative to account for these findings is to recast syntactic complexity in a Relativized Minimality framework (cf. §1.3.1), as proposed by some recent works both in acquisition (Friedmann *et al.*, 2009) and agrammatism (Grillo, 2008). Under this approach, what makes the computation of an object A'-dependency particularly hard to carry out for the human processor is the degree of featural similarity between the two nominal expressions realizing the target (i.e. the moved object) and the intervening element (i.e. the embedded subject).

[44] But the same holds for sentences involving a similar structure, such as object *wh*-questions.

Such a similarity is assessed, in turn, by computing the set of morpho-syntactic features shared by the target and the intervener (e.g. Determiner [+D], common noun [+N], proper name [+Nproper], wh-operator [+Q]), so that the higher the set of (relevant) features shared, the harder the resulting computation[45]. This occurs exactly in (28) and (29), where the target and the intervener belong to the same syntactic category, (cf. (28') and (29')), but crucially not in all the other examples (cf. 30'-32')).

(28') The journalist$_{i\ [+Q,+D,\ +N]}$ who the senator $_{[+D,\ +N]}$ attacked $_{<t_i>}$met the editor.
(29') It was Mark$_{i\ [+Q,+Nprop]}$ who John $_{[+Nprop]}$ attacked $_{<t_i>}$
(30') The journalist$_{i\ [+Q\ +D,\ +N]}$ who Rachel $_{[+Nprop]}$ attacked $_{<t_i>}$ met the editor.
(31') The journalist$_{i\ [+Q,\ +D,\ +N]}$ who I $_{[+D]}$ attacked $_{<t_i>}$ met the editor.
(32') The one$_{i\ [+Q,\ +D]}$ who the senator $_{[+D,\ +N]}$ attacked $_{<t_i>}$ met the editor.

If we accept the featural approach in terms of "intervention" to be on the right track, we can also explain the preference demonstrated by adults in elicited production tasks to avoid the production of standard (active) object relative clauses, such as those displaying the more demanding configuration in (28'), and to rely instead on other structures, compatible in meaning but easier to process, typically the so-called «passive object relatives» (cf. § 1.3.3)[46].

To conclude this section, it seems that the notion of "intervention", reformulated in a locality-based framework, is capable of providing a very subtle metric of syntactic complexity, at least as far as experimental sentence comprehension is concerned. Thus it might be interesting to verify its contribution in measuring the difficulty of written texts, also with respect to readability assessment applications. We will come back to this point in Chapter 3 (§3.7) by discussing a sample of different typologies of relative clauses extracted from the two versions of the corpus of bureaucratic texts under examination.

[45] Cf. Chesi and Moro (2013) and Belletti and Rizzi (2013) for in-depth discussion and a clarification of what makes a morpho-syntactic feature "relevant" for triggering intervention effect.
[46] Without entering into detail on the derivation of these structures (the interested reader is referred to Belletti and Contemori, 2010), it is sufficient to say here that in a passive object relative the intervention effect is not only reduced, as it occurs in (30), (31) and (32), but completely avoided thanks to a preliminary syntactic movement of the whole verbal chunk hosting the internal object.

1.4 Complexity at the level of discourse processing

Up to this point our attempts to characterize the construct of linguistic complexity have remained confined within the sentence boundaries. Yet the process of understanding words and sentences in isolation is different from understanding sentences as part of a discourse. At a deeper level of language comprehension, the reader (or speaker) is also required to assemble each single sentence into a coherent mental representation of the text.

According to an influential theory of discourse processing developed Van Dijk and Kintsch (1983) and Kintsch (1988), readers build three different, and hierarchically arranged, mental representations of the text: first, a verbatim representation that preserves the exact word and syntax of clauses, next a semantic representation (also called *textbase*) describing the meaning of the text in terms of an interrelated network of propositions and, at a deeper and more abstract level, a situation representation (i.e. the so-called *situation model*), which allows the reader to understand the text as an instance of a situation (or a microworld) already established in his/her long-term memory.

The fundamental assumption underlying this theory is that the process of comprehension is highly interactive: it is the ability of the readers to draw inferences from the text, to rely on their prior knowledge in order to retrieve hidden or missing information and to activate existing knowledge structures – variously called *schemata*, *frames*, *scripts* (cfr. Van Dijk and Kintsch, 1983) – which can modulate the ease whereby the situation model will be created and, ultimately, impact on the final comprehension. Under this assumption, a text will be more or less difficult to comprehend to the extent its structure is coherent and fulfils the expectations the reader has incrementally generated, given his/her attentional state, background information, activation of knowledge schemes, etc.

While from this "global" perspective coherence is mainly a psychological construct, which is very hard to be modelled, it also has a counterpart at the level of linguistic structure, which is more directly involved in building the *textbase* representation.

Coherence at this level is generally referred to in terms of *cohesion*, a property of a text that is conveyed by the use of signalling linguistic devices (such as reference, ellipsis, argument overlap, theme-rheme structure) [47], which make explicit the logical links between different units in the texts, so that to minimize the cost of active processing.

[47] Cf. Halliday and Hasan (1976), for some familiar taxonomies of cohesive cues.

This happens, e.g., for the conceptual relation of the type "Problem-Solution", which has been found to be processed faster if linguistically marked by a connective like "because", "as", "since" (Cf. Sanders and Noordman, 2000).

A particular interest in the literature has been devoted to deepening the role of coreference in text processing: this is the property by which two linguistic expressions (an *antecedent* and an *anaphora*) turn out to refer to the same semantic entity in the discourse model. In the next paragraph we will focus on one of the most influential theory of discourse processing dealing with referring expressions, the «centering theory» (Grosz *et al.*, 1983, 1995), by showing its implications in characterizing linguistic complexity at higher levels of processing.

1.4.1 The «centering» model of coherence

The «centering theory» (Grosz *et al.*, 1983, 1995) outlines a model of local coherence, i.e. coherence among the utterances in a discourse segment, which is explained as the result of the interaction between the attentional state of the discourse's participants (listeners or readers) and the different processing demands posed upon them by specific types of referring expressions. The "core" of this approach is the concept of 'centers', which have to be taken as the primitive semantic entities around which the discourse segment unfolds. Centers are distinguished into two categories, called 'forward-looking centers' and 'background-looking center', both of which are linguistically realized by noun phrases (NPs).

According to this model, the initial utterance of a segment contains an order set of 'forward-looking centers' (Cfs), i.e. the potential "foci of attention" to which the subsequent utterance can refer. Such a hierarchy is determined by a combination of syntactic, semantic and discourse properties: in particular, for what concerns syntactic information, the grammatical role of each center plays a major importance, so that the order [subject > object > others][48] is taken as the preferential way to establish the NPs salience. Each utterance other than the segment initial one realizes only a single entity, i.e. the 'background-looking center' (Cb), which provides the link to the preceding utterance.

Within locally coherent discourses, the Cb corresponds to the highest ranked Cf and it tends to be preserved across pairs of transitions between consecutive utterances, that is to say that sequences of Cb's continuation are preferred over Cb's shifts.

[48] A further distinction was then established between direct and indirect object.

To make these points clearer, let's consider two different ways of conveying the same informational content in discourse (33) and (34) (for each utterance, Cb and Cfs are indicated in brackets)[49].

(33) a. John went to his favourite music store to buy a piano. (Cf = {John, the store, a piano})
b. He had frequented the store for many years. (Cb = John; Cf = {John, the store, years})
c. He was excited that he could finally buy a piano. (Cb = John; Cf = {John, a piano})

b. He arrived just as the store was closing for the day (Cb = John; Cf = {John, the store, the day})

(34) a. John went to his favourite music store to buy a piano. (Cf = {John, the store, a piano})
b. It was a store John had frequented for many years. (Cb = the store; Cf = {the store, John, years})
c. He was excited that he could finally buy a piano. (Cb = John; Cf = {John, a piano})
d. It was closing just as John arrived. (Cb = the store; Cf = {the store, John})

As we can see, the sequence of sentences reported in (33) is intuitively more coherent than the one in (34): this is because the reader is able to identify in a clearer manner that the center of attention is "John", which is kept constant throughout the segment. The same does not hold in (34), whose structure makes it much more difficult to decide whether the focus is John, the store, or his desire to buy a piano.

This example provides evidence that each utterance within a discourse segment has a unique focus that is more prominent and this focus models the reader's attention and the expectations about what is to come. To the extent these expectations are welcomed, a discourse is perceived as more coherent, thus easier to comprehend, because fewer inferences are required by the reader.

A qualifying point of the *centering* framework deals with the linguistic realization of the Cb. Given a certain utterance, the probability of an entity to realize the Cb in the current utterance is determined by the linguistic realization of the noun phrase denoting the entity, and specifically by the choice of the referring expression. At this level, a major distinction is established between pronouns and full referring expressions (e.g. nouns).

[49] The example is taken from (Grosz *et al.*, 1995).

While the former are typically used to refer to entities already mentioned in the discourse, the latter introduce new entities in the discourse[50]; thus, for the purpose of coherence, pronouns are the prototypical cue to establish the Cb, whereas full NPs are more functional to signal Cb transitions[51].

The *centering* predictions about the role of referring expressions in discourse processing have received empirical support in reading comprehension's research. In particular, Gordon *et al.* (1993) and Kennison and Gordon (1997) reported a higher reading latency[52] for sentences containing a repeated name rather than a pronoun, which was greater when the antecedent introducing the referent in the preceding sentence was syntactically prominent. This effect, the so-called *repeated-name penalty*, was observed when comparing the reading times for sentences such as (35c) and (35c'), presented after the sequence of utterances (35a-b).

(35) (a) Susan really likes animals.
(b) The other day she gave Betsy a pet hamster.
(c) **She** reminded her/Betsy that such hamsters are quite shy and need gentle handling.
(c') **Susan** reminded her/Betsy that such hamsters are quite shy and need gentle handling.

On the contrary, no detrimental effect was assessed using the same noun for direct objects, that is to say that the realization of the direct object either as a repeated noun or a pronoun was ineffective with respect to reading ease/difficulty.

Besides, the *repeated-name penalty* was also reduced for sentences like (36c) within the discourse (36), in which the repeated name denotes an entity that was not prominent in the previous sentence.

(36) (a) Sue knew Tom wanted the St. Bernard puppy in the store.
(b) She offered to buy it for him as a Christmas present.

[50] See Gordon and Hendrick (1998), Gordon *et al.* (2001) for a formal analysis accounting for the discourse properties licensing the use of reduced vs. full referring expressions.
[51] Note that the centering approach does not assume that the use of pronoun is the only possible way to realize the Cb, yet such a condition has to be ensured if any of the less highly ranked forward-looking centers has been realized as a pronoun in the same utterance.
[52] Both these studies reached the same conclusions by adopting a different experimental paradigm: the self-paced reading in (Gordon *et al*., 1993) and the eye-tracking in (Kennison and Gordon, 1998).

(c) **Tom** said that St. Bernard generally makes wonderful present.

(c') **He** said that St. Bernard generally makes wonderful present.

According to the *centering* framework, such an example triggers a *center shift* condition, for which, indeed, it is predicted the less advantage of using pronouns for the purpose of coherence.

1.5 Summary

The intent of this chapter was to provide an overview of the notion of linguistic complexity from the point of view of human language processing. As we tried to highlight, this concept is an 'umbrella' term and covers a wide assortment of properties digging into the multi-layered structure of language domains, thus making difficult, if not impossible, to define a unique metric of complexity. A particular emphasis has been dedicated to syntactic processing, as we believe – following Scott (2009: 184) – that «if a reader cannot derive meaning from individual sentences that make up a text, that is going to be a major obstacle in text-level comprehension».

It has to be pointed out that, although many of the properties discussed in these pages are very subtle and we do not expect them to affect normal offline reading to the same extent they manifest in online sentence comprehension, the same might not hold for "atypical" populations. It is the case of language-impaired readers, e.g. aphasics, who might struggle in capturing the meaning of a text because of the presence of complex sentences displaying canonicity effects (§1.3.3), or dyslexic readers, for whom the advantage deriving from controlling specific lexical variables, such as word length or orthographic neighbours size (§1.2), might leave far available resources for reading processes following word encoding.

On one side, such considerations point out the importance of conceiving flexible, i.e. tuned to the needs of the final readership, metrics of linguistic complexity; on the other side, they also raise the question of whether and to what degree of approximation, these metrics can be operationalized by adopting a NLP-based perspective to text difficulty analysis.

Next chapter addresses exactly these issues by focusing on a specific field of NLP research in which the operationalization of factors involved in text complexity is a methodological precondition: the automatic assessment of text readability.

Chapter 2

Operationalizing linguistic complexity from a NLP perspective: the computational assessment of text readability

2.1 Readability and readability assessment: a disputed topic

Automatic readability assessment is typically defined as the task aimed at providing an objective and quantifiable prediction of how difficult a text is to read and understand, by investigating its linguistic structure. From this very broad definition, two implications seem to derive: i) that text complexity can be measured by means of unbiased formulae and ii) that having a measure of readability is predictive of understanding, or saying it differently, that the comprehension process is the outcome of a readable text.

Traditionally, indeed, this task has been pursued by means of the so-called "readability formulae", namely mathematical equations that compute certain constants and a few (usually not more than two) parameters taken from the text, in order to yield a readability score for the text under examination. As we will see in the next paragraph, a functional perspective – sometimes referred to as the «positivist paradigm» (Wray and Janan, 2013) – led the origin of these formulae: providing educators with a simple method to select more suitable materials for students, according to their level of reading. Consequently, the scores of the most popular formulae have been generally "qualified" by calculating statistical correlations between the observable textual features and the reading levels of readers, as measured by standardized tests (e.g., the cloze test)[53]. However, things are far from being so clear-cut and the reason lies in the object of evaluation itself: readability, indeed, continues to be «among the most discussed, misunderstood and misused concept in reading»[54].

[53] Cloze test involves the systematic deletion of words in a given text, which the examined subjects are required to retrieve on the basis of the context, by generating them or by selecting the word from a set of options. Text comprehension is measured by how accurately the reader can fill in the blanks with an appropriate word. This technique was first discovered by the German scholar Ebbinghaus in 1897 and became widely used as a tool for measuring readability, although the method has been widely debated.
[54] Pikulski, J.J., 2002. *Readability*. U.S.A: Houghton Mifflin Company, p.1.

With this respect, a first distinction needs to be done between readability and legibility. The latter has been defined as «the ability to recognize a letter, number, or word easily and correctly outside of the context of the word, sentence, or phrase» (Hasbrouck, Tindal, Parker, 1994) and it stems from characteristics that are generally related to the graphical aspect of the text, such as typeface, layout, contrast, edge sharpeness, among many others. Although legibility is one of the variables of a readable text, and it can greatly impact on the reading performance of specific categories of readers, e.g. dyslexics (Rello and Baeza-Yates, 2013), the concept of readability requires a broader view to the potential sources from which text complexity might descend, which can be classified into two general classes along the linguistic/extra-linguistic distinction.

The former includes all those features that are identifiable by looking inside the linguistic realm of a text, thus lexical, syntactic, semantic and discourse-related features, as well as properties affecting the style and textual genre. In the latter we can ascribe reader-oriented variables, such as her/his degree of familiarity with the topic, normal or impaired cognitive development, proficiency's level in the language used in the text to be read (e.g. L1 vs. L2), interests and motivations (cf. Wray and Janan, 2013 for a review). There is indeed a large consensus in psychological literature to consider reading an interactive process (Rumelhart, 1977; Just and Carpenter, 1980;), in which the reader is not passive but instead actively involved; consequently, readability assessment should aim to «effect a 'best match' between readers and texts […] thus optimal difficulty comes from an interaction among the text, the reader and his/her purpose for reading», as claimed by two prominent scholars of the field, Jeanne Chall and Edgar Dale (1995: 45-46).

The multifaceted and interactive character of the concept of readability makes it rather impossible to envisage an absolute and objective way of measuring it, thus casting doubts on the potentiality itself of a mathematical formula to embody that «best match» advocated by Dale and Chall. This does not lead to reject the original assumption underlying the traditional approach to automatic readability assessment, i.e. the possibility of getting a reliable approximation of the difficulties a reader might encounter in a text by measuring linguistic predictors of text complexity. Such an intuition seems theoretically well justified: as we revised in Chapter 1, linguistic theory has nowadays reached a point where formal explanations and empirical results can succeed in making finer predictions about the properties of linguistic objects that may have a real impact on comprehension to varying degrees and at different stages of processing.

Nevertheless, for a real progress in automated readability assessment research, it was necessary not only a theoretical refinement of the notion of linguistic complexity, but also the possibility of making such a notion computationally manageable. This is the topic of the present chapter, which is organized as follows: section 2.2 reviews the traditional approach to automatic readability assessment, while section 2.3 focuses on the current perspective developed in the NLP community; in particular, paragraph 2.4 introduces READ-IT, which is the only existing "advanced" tool for the automatic readability assessment of Italian texts. The overview of the linguistic features underlying this tool, and the way they can be extracted and monitored from automatically parsed texts, will also provide the necessary methodological "background" to understand the linguistic profiling investigation discussed in Chapter 3.

2.2 Classic approach to automatic readability assessment: the readability formulae

Research devoted to making text readability an objective measure traced back to early last century, when readability formulae started being investigated in the field of U.S. education. A major concern was indeed to endow teachers and publishers with a simple tool to select schoolbooks according to the reading's level of their audience. This was particularly crucial in light of the increased percentage of first-generation immigrants coming into schools, who faced difficulties in comprehending standard educational materials[55].

A first attempt in this direction was the publication of the *Teacher's Word Book* in 1921 by the psychologist Edward L. Thorndike. Based on a survey of hundreds of texts from different sources (children's novels, English classics, schoolbooks, hobby literature), Thorndike came up with an alphabetical list of 10,000 words enumerated by frequency of use. Such a resource, very innovative at that time, was intended to serve as a scientific tool to measure educational materials' difficulty, under the assumption that word frequency was also directly correlated to the reader's familiarity and ease of use.

In this respect, the Thorndike's list led two scholars, Mabel Vogel and Carleton Washburne, to develop the first readability formula based on methods from statistical linguistics and empirical evaluation: the *Winnetka formula* (1928)[56].

[55] Cf. DuBay, William, H. (2006). *The classic readability studies*, Impact Information Costa Mesa.
[56] Vogel, M. and Washburne, C. (1928). *An objective method of determining grade placement of children's reading material*. Elementary School Journal, 28, 373-381, In Dubay, *The classic readability studies*, pp. 18-26.

The formula was created on a sample of 700 books reported by 37,000 children as books they had read and enjoyed. The books were measured along several linguistic features, covering both lexical and morpho-syntactic aspects, one of which was exactly the number of words not contained in the Thorndike's list. Each feature was then correlated with the mean reading level of the children participating in the survey, as assessed by the Stanford Achievement test on paragraph sections taken from the examined texts. The features that combined together yielded the best multiple correlation score[57] (equal to 0.845), entered into a regression equation, predicting «with a high degree of reliability the reading score necessary for the reading and understanding of any given book»[58].

Similar to this formula, but specifically conceived for adults with limited reading skills, was the one created by William S. Gray and Bernice Leary in 1935, under the ambitious title *What makes a book readable*. Their study was indeed the first, and for many years the only one, to carry out a comprehensive linguistic investigation covering more than 200 elements affecting reading ease, which were grouped into four categories (content, style, format, features of organization). Among the whole examined features, the authors focused on the subset of 48 stylistic variables, as they were the most amenable to reliable quantitative measurement and verification. As in the case of the *Winnetka formula*, such features were then correlated with the scores of the reading-comprehension tests performed by more than 800 adults on the material under examination. The higher level of correlation (i.e. 0.645) was achieved by combining five variables[59], which were used to create their own index.

Since these pioneering studies the rate of new formulae rapidly increased, as well as the interest they aroused within a variety of fields, such as government publications, journalism and business communication, military and health information. Nevertheless, their underpinning assumptions remained almost unchanged and the researchers focused on simplifying the algorithms by minimizing the number of features, so that to overcome the efforts of manual counts. In the majority of cases, the best predictive power was reached by combining only two measures, accounting respectively for a lexical and a syntactic dimension of text complexity.

[57] These features were the following: number of different words in a 1000-word sampling; total number of prepositions in a 1000-word sampling; total number of uncommon words, *i.e.* not contained in the Thorndike's list; number of simple sentences in 75-sentence sampling.
[58] *Ivi.*, p. 21
[59] Namely: average sentence length in words; number of different "hard" words (not contained in a list of familiar words); number of first, second and third-person pronouns; percentage of different words; percentage of prepositional phrases.

In particular, if sentence length became the typical proxy of syntactic difficulty, several other variables were proposed to assess the vocabulary load posed on the reader. Some of the formulae continued to consider word familiarity (derived by some lists of frequency) as the strongest predictor of readability: it is the case of the *Dall-Chall formula*, probably the most famous one on the side of vocabulary-based index.

This formula, originally proposed in 1948, made reference to a list containing the 763 words resulted familiar to the 80% of fourth-grade readers[60], as assessed by empirical tests. The percentage of "hard words" (i.e. words outside the list), added to the average sentence length for 100-word samples, produced the most consistent correlation with the reading test scores (.70).

Similar results were obtained by replacing the percentage of unfamiliar words with the average word length, as firstly demonstrated by Rudolph Flesh in his 1948 famous work. The Flesh Formula was originally composed by two parts: one for assessing the reading ease and the second to predict "human interest". The first part (i.e. the "Flesh Reading Ease", cf. table 1) was obtained by calculating the average sentence length, in terms of different words, and the average word length, in terms of syllables, for samples of 100 words. These values, subtracted from a statistical constant, allowed the final score to range on a scale from 0 to 100, where 0 indicates the minimum and 100 the maximum readability.

Reading Ease = 206.835 – (1.015 X ASL) – (84.6 X ASW) ASL = average sentence length ASW = average number of syllables per word

Table 1: Flesh Reading Ease Formula.

For what concerns the prediction about the interest that a text could arouse in the reader, the second part of the formula took into account the number of the so-called *personal words* and *personal sentences*. The former were defined as «all nouns with natural gender; all pronouns except neuter pronouns, and the words *people* (used with the plural verb) and *folks*», while personal sentences identified «spoken sentences, marked by quotation marks or otherwise; questions, commands, requests, and other sentences directly addressed to the reader;

[60] In 1995 the formula was revised by the same authors expanding the list to 3,000 words.

exclamation; and grammatically incomplete sentences whose meaning has to be inferred from the context».[61]

Human Interest = 3.635 pw + 314 ps Pw= personal words Ps= personal sentences

Table 2: Flesh Human Interest Formula.

However, the equation for human interest prediction did not gain large consensus, because it failed to reach high correlations with the scores of the standard reading tests which, as we pointed out, were still the most confident validity metric. On the contrary, the Flesh Reading Ease formula became widely used and was officially adopted after its 1976 revision[62] for grading educational materials by U.S. school system.

It has to be noticed that both sentence and word length not only could be easily derived by using shallow text analysis tools but, to a certain extent, their implication in reading had received empirical validation. For instance, Edward Zipf's research in statistical linguistics demonstrated the correlation between word length and word frequency (known as "first Zip's law", 1935), according to which the longer the word, the less frequent it is within corpora[63]. Thus, on the assumption that word frequency was the main factor involved in vocabulary ease, calculating word length could offer an effective mean and reduce the need of relying on external resources (e.g. the list of familiar words), which were certainly more qualitative, yet time-consuming and variable across readership and domains.

This is also the approach characterizing *Gulpease*[64], the first and most famous readability "traditional" index specifically conceived for Italian, which is briefly introduced in the next paragraph.

[61] Flesh R., *A New Readability Yardstick*, Journal of Applied Psychology, Vol. 32(3), Jun 1948, 221-233.

[62] The revision was conducted by J. Peter Kincaid for a project commissioned by the US navy. It led to a new formula, known as the Flesh-Kincaid Grade Level, which converted the original Flesh Formula scores into grade levels.

[63] However, it is worth pointing out that the shorter the word, the more ambiguous it is, hence words appearing frequently within different corpora might be familiar in one but not all of their meanings. For what concerns Italian, e.g., if we inspect the first ten most frequent nouns from the Lessico di frequenza dell'italiano parlato (LIP) and the Lessico di frequenza della lingua italiana contemporanea (LIF) (data reported in Voghera (2001)), we observe that in both the corpora the nouns with more than 10 meanings are all disyllabic.

[64] Lucisano, P., Piemontese, M. E. (1988), *Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana*. In: «Scuola e città», 34, pp. 110-124.

2.2.1 The *Gulpease Index* for Italian

As we mentioned in the previous paragraph, the quantitative approach to readability assessment developed in U.S. and the earlier formulae becoming popular were based on English language. It is only by the end of 1970s that research on readability and its automatic treatment spread to other countries, so that we can find studies for Finnish, French, Danish, Swedish and Dutch[65].

For what concerns the Italian language, an initial attempt towards a language-specific formula was made by Roberto Vacca, who developed a first formula in 1972 and revised it in 1986. Both the original and the revised one were merely an adaptation of the Flesh formula, as the only modification affected the value of the coefficient assigned to the two variables (i.e. sentence length and word length).

Instead, the first formula specifically tailored for our language was the *Gulpease index* – named after its creators[66] – and still quite popular in the Italian context. Basically, the development of this formula followed the traditional approach to readability assessment, in that: i) the difficulty of a sample of texts (taken from authentic school textbooks) was evaluated by means of comprehension tests targeting different populations (i.e., Italian students enrolled in the last year of each school cycle[67]); ii) the examined texts were measured across different linguistic variables; iii) the correlations between the levels of difficulty and the linguistic variables were statistically calculated; iiii) the linguistic variables with a higher correlation were selected to enter into a regression equation.

As for the classic English-based formulae, the most reliable predictors turned out to be sentence length and word length, although the latter was here calculated in terms of number of characters rather than syllables[68].

[65] *Ibid*., p. 112.

[66]The acronym Gulp stands for "Gruppo Universitario Linguistico Pedagogico" (Linguistic Pedagogical Academic Group), the research group composed by linguists, education specialists and computer scientists, in charge of the project.

[67] The fifth year of the primary school, the third of the lower secondary school and the fifth of the upper secondary school, in accordance with Italian education system.

[68] Other combinations of variables were also tested in this work: the authors reported that the so-called index *Gulpsynt* - based on the percentage of fundamental words (i.e. content words belonging to a first version of the *Basic Italian Vocabulary*, De Mauro 1980), number of words for 100-word samples and percentage of subordinate conjunctions out of the total of words - turned out as being the most correlated one with the comprehension scores; however, it was discarded because similar correlations could have been achieved by using the Gulpease index, which was more practically manageable by first calculators.

With respect to the text readability score, the Gulpease index ranges from 0 to 100 indicating, respectively, the lowest and the highest value of readability. The final quantitative value has been also mapped with respect to "categories" of readers, which were defined on the basis of the comprehension scores obtained by the three empirical samples (i.e. elementary, junior high school, senior high school), so that to make it possible a more qualitative interpretation of the index.

Just to give an example, a text achieving a score of 50 has to be taken as fairly accessible to readers with a senior high school diploma, quite difficult for readers with a junior high school diploma and very difficult to comprehend for people with an elementary school diploma.

Table 3 illustrates the formula and its interpretative value.

$$Gulpease = 89 - LP/10 + 3\,FR$$

LP = characters/100 words
F = n° sentences/100 words

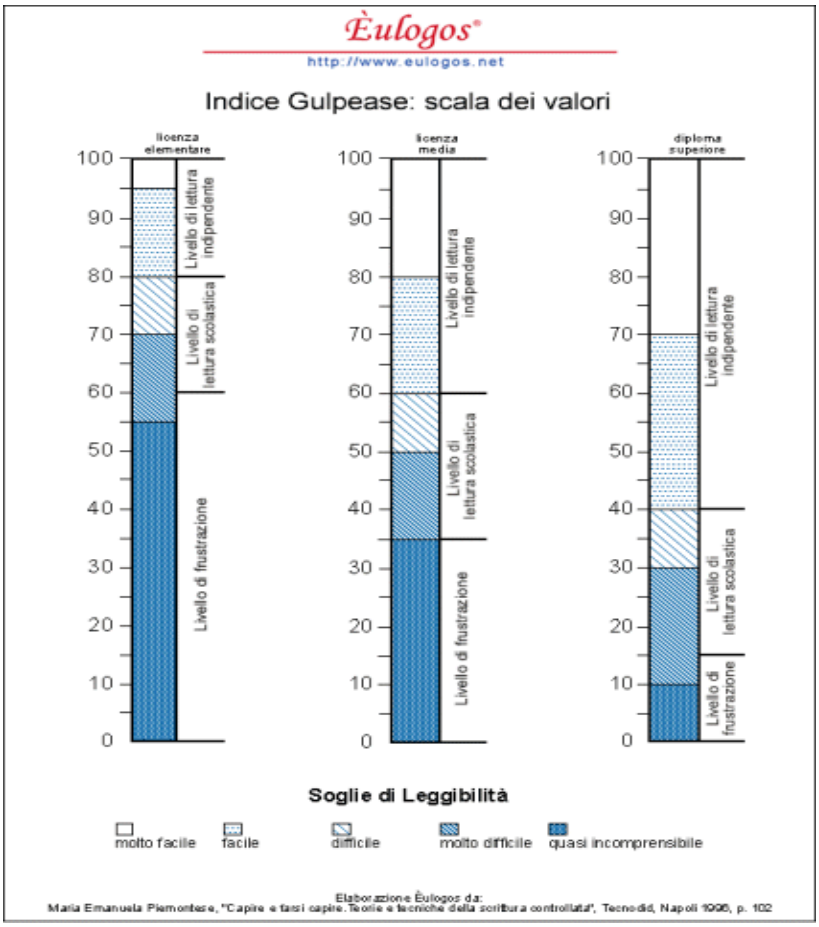


Table 3: The *Gulpease* Index.

2.3 New generation of automatic readability indexes

While research in linguistics and psycholinguistics within the domain of cognitive sciences has provided a better understanding of the mechanisms underlying language processing, it had the consequence to quit readability investigation for almost 15-20 years. The *positivist* approach embodied by the readability formulae was clearly unsatisfactory to give a real approximation of the difficulties a text might pose to a reader and underwent severe criticism (see, among many others, Davison and Kantor, 1982).

Some studies targeting young students' comprehension demonstrated that texts rewritten according to the requirements of traditional formulae (that is, reducing sentence length and preferring short words to longer ones) did not improve comprehension (Green and Olsen, 1986)[69]. Similarly, the rewriting of highly technical texts along the indications of the Flesh Formula, proved to be ineffective to help non-experts readers to recall the main concepts they read (Charrow, 1988)[70].

It has to be noted that the "poor" sophistication of traditional readability formulae, being so far from assessing cognitive proxies of linguistic complexity, was already known to their creators; yet, they were constrained by the limits of the existing text analysis tools, which did not allow for a computational measurement of textual complexity predictors other than the surface-level ones (typically sentence length and word length). However, over the last two decades, the higher accuracy of automatic linguistic annotation and machine learning techniques for text analysis, as well as the development of machine-readable lexicons containing a wealth of linguistic information (e.g. WordNet, §2.3.1.1), have made it possible to overcome many of these limitations. We are thus experiencing a renewed interest for readability assessment technologies, with the emergence of a new generation of readability indexes, improved in their explanatory power and linguistic plausibility.

As claimed by François and Fairon (2012), whose work addresses automatic text readability for French: «three main ingredients characterize these new formulae: the use of a large number of texts assessed by experts (coming from textbooks, simplified newspapers or web resources) as training data; the use of NLP-enabled features able to capture a wider range of readability factors and the combination of these features through a machine learning algorithm».

[69] In Davison and Green (1988), pp. 115-140;
[70] *Ivi,* pp. 85-114;

We believe that a fourth element should be added to this list, that is the customizable nature of these indexes: indeed, the selection of the linguistic features, which stands at the core of the more recent algorithms, has been shown to vary with respect to the particular audience they are intended to, the applicative viewpoint from which readability assessment is carried out and the typologies of texts (e.g. general or genre-specific) to be covered.

In the paragraphs that follow, we will provide a more detailed description of the current approach to automatic readability assessment research. Rather than presenting a comprehensive summary of the literature (cf. the update survey reported in Collins-Thompson, 2014), a selection of some recent works will be offered, with the aim of explaining what the concept of NLP-enabled features stands for, how these features succeed in approximating some linguistic markers of complexity described in Chapter 1, as well as the requirements (in terms of formalisms, resources and methods) that enable their measurement.

2.3.1 Lexical features

The evaluation of more fine-grained attributes involved in lexical complexity, such as those described in paragraph 1.2, has been promoted by the availability of machine-usable dictionaries enriched with such specific attributes. In what follows we describe some of these major resources and illustrate how they have been integrated into new readability indexes.

2.3.1.1 WordNet

Originally developed for English language at the Princeton University Department of Psychology starting from 1984, WordNet (Fellbaum, 1998; Miller, 1990) is now available for more than forty languages[71].

It is a lexical database inspired by current psycholinguistic principles on human lexical memory, according to which the mental lexicon is structured into clusters of semantically related words (cf. § 1.2). Following such indications, the lexical entries contained in WordNet (which are distinguished into nouns, verbs, adjectives, adverbs) are organized into sets of cognitive synonyms (called *synsets*), each one representing the underlying lexical concept.

[71] An updated list can be found at: http://globalwordnet.org/wordnets-in-the-world/ [last access: 01/07/2015]

Synsets are linked each other by means of lexical and semantic relations, which try to mirror the way speakers organize their own mental lexicon. These relations include hypernymy/hyponymy (i.e. the relation of being subordinate or belonging to a lower rank or class, also known as ISA), antonymy, entailment (logical inference), and meronymy/holonymy (i.e. the relation that holds between a part and the whole).

Without entering into further details on the database structure and underlying tenets[72], we now focus on the use of WordNet in computational readability assessment research. With this respect, a well-designed proposal was advanced by Lin and colleagues in 2009, who relied on this tool with the aim of deriving an index of lexical complexity, called *basic level noun ratio*. Moving in the framework of the *Prototype Theory* (cf. § 1.2), the authors assumed that the higher the value of this index, the more readable the resulting text. With this respect, it has to be noted that such a piece of information, although cognitively salient, is not directly encoded in WordNet. Thus, in order to understand which nouns might fall in the basic level category, the authors hypothesized that a basic level word is likely to be shorter and morphophonemic than its hypernyms and hyponyms. This assumption was tested on the 18 basic level nouns originally identified by Rosch (1976): for each of these nouns, the authors extracted the corresponding taxonomy in WordNet and assessed the following parameters:

1) the ratio of compounds containing the target word in its full list of hyponyms;
2) the length difference of the noun (in terms of letters) with respect to the average length of its hyponyms.

This inspection turned out to confirm that the basic level nouns were not only shorter than the average length of both their hypernyms and hyponyms, but also more frequently used to create compounds with respect to the latter. Based on these findings, the authors formulated a *Filter Condition*, in which the compound ratio (*i.e.* the number of the hyponyms containing the target word divided by the number of the target word's full hyponyms/hypernyms) was set at $\geq 20\%$ and the length difference (*i.e.* the average length of the target word's full hyponyms minus the length of the target word) at ≥ 2. The nouns passing the Filter Condition were taken as being basic level nouns and the ratio between the basic level nouns and the total nouns in a text gave the *basic level noun index*.

[72] The interested reader can consult the official website for extensive bibliography, at the following link: http://WordNet.princeton.edu

The last step consisted in evaluating the likelihood of this index to predict text readability. In this regard, two corpora of online graded readings (i.e. English readings for children vs. for high school students) were selected and the index was calculated for all texts, along with the scores provided by a wide range of more traditional readability formulae. In line with the expectations, the grade level and the basic level noun ratio turned out to be inversely correlated, showing that the values reported by the index diminished progressively across the grade-level progression. On the contrary, the predictive power of the classic readability formulae was much less reliable and consistent across graded school texts.

Although in a more naive way, WordNet has also been adopted by Graesser *et al.* (2004) for the development of *Coh-Metrix*[73], a web-based software tool that, rather than yielding a unique readability score for an input text, evaluates it on more than 200 features selected as proxies of cohesion and text difficulty at various levels of linguistic, discourse and conceptual structure. These features are distinguishable into five classes: general word and text information; syntactic indices; referential and semantic indexes; indexes for situation model dimensions and standard readability indexes. Among the features of the first class, WordNet is exploited to evaluate word ambiguity and word abstractness metrics, which are assessed by averaging, respectively, the number of synsets and the number of hypernym levels for each word of the text having a correspondent entry in the lexical dataset. The number of synsets in which a word appears (*i.e.* the polisemy value) is considered as an indirect measure of its ambiguity, while the number of hierarchical levels intervening between the synset containing the lemma and its upper hypernym provides a metric of abstractness.

It is worth noting that Coh-Metrix has inspired several adaptations for other languages, such as the *Coh-Metrix-PORT tool* (Scarton *et al.* 2009; *Aluísio et al.*, 2010) for Brazilian Portuguese. A recent version has also been realized for the Italian language by Tonelli *et al.* (2012)[74], where the features so far discussed have been measured against MultiWordNet (Pianta *et al.*, 2002), the Italian aligned version of the original WordNet.

2.3.1.2 Medical Resource Council (MRC) Psycholinguistic Database

The Medical Resource Council (MRC) Psycholinguistic Database (Coltheart, 1981) is another machine usable dictionary that provides a large number of lexical variables inspired

[73] http://co-metrix.memphis.edu
[74] The web interface is available at: http://terence.fbk.eu/services/api/computeReadability/v2/

by word processing research. It was originally conceived to provide balanced stimuli for psychometrics tests, but also to support researchers in Artificial Intelligence and computer scientists in the design of text processors and NLP tools.

Currently, the MRC Psycholinguistic Database, updated in 1988, contains 150,837 English words and accounts for about 26 different linguistic properties, although some of them are not available for the whole set of entries[75].

As it can be seen in table 4, there are several features involved in defining word complexity (§1.2), justifying its adoption in the context of automatic readability assessment of texts. This is what has been done, again, in Coh-Metrix (cf. the previous paragraph), which provides a measure of vocabulary difficulty based on the following six MRC properties: familiarity, concreteness, imageability, colorado meaningfulness, paivio meaningfulness, age of acquisition.

	Name	Property	Occurrences
1	NLET	Number of letters in the word	150,837
2	NPHON	Number of phonemes in the word	38,438
3	NSYL	Number of syllables in the word	89,402
4	K-F-FREQ	Kučera and Francis written frequency	29,778
5	K-F-NCATS	Kučera and Francis number of categories	29,778
6	K-F-NSAMP	Kučera and Francis number of samples	29,778
7	T-L-FREQ	Thorndike-Lorge frequency	25,308
8	BROWN-FREQ	Brown verbal frequency	14,529
9	FAM	Familiarity	9,392
10	CONC	Concreteness	8,228
11	IMAG	Imagery	9,240
12	MEANC	Mean Colorado meaningfulness	5,450
13	MEANP	Mean Paivio meaningfulness	1,504
14	AOA	Age of acquisition	3,503
15	TQ2	Type	44,976
16	WTYPE	Part of speech	150,769
17	PDWTYPE	PD part of speech	38,390
18	ALPHSYL	Alphasyllable	15,938
19	STATUS	Status	89,550
20	VAR	Variant phoneme	1,445
21	CAP	Written capitalised	4,585
22	IRREG	Irregular plural	23,111
23	WORD	The actual word	150,837
24	PHON	Phonetic transcription	38,420
25	DPHON	Edited phonetic transcription	136,982
26	STRESS	Stress pattern	38,390

Table 4: MCR dictionary linguistic properties.

[75] The source data of the MRC Psycholinguistic Database can be downloaded from the official page: http://www.psych.rl.ac.uk/ [last access: 01/07/2015]

It goes without saying that less-resourced languages (e.g. Italian, as far as readability assessment is concerned) have to cope with the unavailability, or poorer coverage, of such detailed lexical databases[76]; this makes it rather necessary to supply them with more traditional resources, such as word frequency and word familiarity lists drawn from large corpora (e.g. the Basic Vocabulary for the Italian Language, recently updated in De Mauro and Chiari, 2014).

2.3.2 Semantic and discourse related features

As noted in section 1.4, the comprehension of written texts is also affected by high-level cognitive properties, i.e. those properties assuming relevance at deeper levels of processing, when the reader has to connect each separate sentence into a unique coherent representation.

The computational treatment of discourse features, such as cohesion and coherence, for text readability tasks is still in a preliminary stage, yet it is possible to find some recent works which have addressed it, with the aim of adapting readability metrics for the needs of particular categories. It is the case of Feng *et al*. (2009), who devised a new readability metric for adults with intellectual disabilities (ID)[77], in which the incorporation of discourse level features has been prompted by the selective impairments affecting the intended target.

As reported by the authors, ID readers find it much more difficult to build a cohesive model of the discourse rather than decoding single words and sentences. This deficit seems to be related to a limited working memory span, which slows down the semantic encoding of new information; as a consequence, the smallest the set of discourse referents within a text (both at sentence and document level), the more comprehensible the text would be to them. In addition, since the process of generating a meaningful discourse representation not only requires the reader to keep track of the detected entities, but also to resolve the semantic relations between them, a text can be less or more challenging to comprehend to the extent it displays a lower or higher distance between related entities.

[76] For what concerns Italian, Burani *et al*. (2001) developed an electronic lexical database, inspired to psycholinguistic evidence and containing the following variables: age of acquisition, familiarity, imageability, concreteness, adult written frequency, child written frequency, adult spoken frequency, number of orthographic neighbors, mean bigram frequency, length in syllables, and length in letters. Nevertheless, it only covers 626 nouns.

[77] Specifically, those classified in the mild level of mental retardation with an IQ scores between 55 and 70.

To try to model these properties by relying on NLP tools, the authors first defined the concept of entity as a type "person", "location" and "organization", which were extracted by an entity detection module. Based on this output, two typologies of entity-based features were implemented. The first typology accounts for the concept of «entity density» and it was calculated by measuring:

- the total number and the average number of entity mentions (each token corresponding to an entity) at sentence and document level;
- the total number and the average number of "unique" entities (*i.e.* occurrences of the same entity are counted once), at both sentence and document level.

The second typology is meant to manage the concept of «lexical chain» and makes use of a WordNet-based algorithm, which allows for checking whether the entities contained in the document are related by synonym, hypernym/hyponym and sibling relations by looking at the corresponding WordNet entry: if such a condition occurs, a lexical chain in established. For each of this chain, the following measures were then calculated:

- the average length of the lexical chain (in terms of number of entities within the same chain);
- the average lexical chain span (in terms of number of words occurring in the document between the first and the last member of the chain);
- the average number of «active» chains for each word and each noun phrase: this measure calculates, for each word and each entity of the text, the number of lexical chain spans in which the current word/entity is included. According to the authors, the concept of «active» lexical chain intends to capture the «total number of concepts that the reader needs to keep in memory during a specific moment in time when reading a text».

All of these features, in addition to a set of so-called "parse-tree features" (e.g. average parse tree height, average number of noun phrases, verb phrases *etc.*), (cf. next paragraph), were extracted from a training corpus consisting of texts previously labelled according to a predefined readability class[78].

[78] The corpus consists of the following material: a selection of paired (*i.e.* children/adults) documents (drawn from Encyclopedia Britannica and from a collection of local CNN stories) and a selection of texts labeled by school grades (drawn from the Weekly Reader Corpus). Although they are not specifically conceived for the intended audience, it is assumed that both the texts for children and those for lower

To evaluate the effectiveness of the new discourse-based properties for the readability assessment task, the whole set of features was differently combined so that to define three regression models: a first model was trained on parse-tree features only, a second model on the new entity-based features and a third model on both the typologies. For both the evaluation methods (depending on the testing corpus, cf. footnote 78), the model including discourse-based attributes yielded a more accurate prediction of the reading difficulty of the examined texts[79].

While the study so far discussed proposed an effective way to handle discourse properties on the basis of the notion of entity density and lexical chains, other works have tried to simulated a model of coherence through an entity-grid approach (Barzilay and Lapata, 2008; Pitler and Nenkova, 2008; François and Fairon, 2012).

In particular, on the assumption that locally coherent texts exhibit certain regularities which have a counterpart at the linguistic level, some of these models have incorporated the *Centering*'s definitions of focus continuity and focus shifts (cf. §1.4.1) to predict text readability. Lapata and Barzilay (2005) moved the first steps in this direction by training an algorithm to model coherence based on a gold corpus of simplified texts assumed as indicative of a high coherence. For each text of the gold corpus automatically parsed, the algorithm detected the entities and assigned them a grammatical role (e.g. subject (S), object (O), other (X), or none (-), if the entity is not present). All the types of "transitions" for adjacent sentences (out of the 16 potential ones)[80] were then calculated: for example, if an entity was the subject of a previous sentence but the object in the next one, the system recognized a type of 'S-O' transition. The whole pairs of sentence transitions found in the training data provided a model of coherence, which was then used to establish the degree of coherence of a new text. In this case too, the model resulting from the incorporation of these

grades provide an example of easier texts, in terms of internal coherence and inference load; thus, they could be used as a valid benchmark to investigate how a text should appear in order to facilitate the comprehension of ID people. Besides, as part of a feasibility study for empirically testing the level of comprehension of this population, the authors also collected a small corpus of 20 paired (original and manually simplified) newspaper articles (i.e. LocalNews Corpus) and provided each text with the resulting comprehension scores. However, this corpus was primarily used for evaluation and not for training the readability models, because of its small size.

[79] The model derived from the combination of the parse-tree and discourse-related features outperforms the other models in predicting the grade level of the Weekly articles (average error: first model: 0.6032; second model 0.6110; third model: **0.5650**). Even more interesting are the results of the second task, where the system is evaluated against texts labeled with the actual comprehension scores of the ID readers; here the model trained on only the discourse-related features achieves the best correlation scores (R correlation: first model: -0.283; second model: **0.352**; third model: 0.342).

[80] The 16 transition patterns are: "SS", "SO", "SX", "S-", "OS", "OO","OX","O-", "XS", "XO", "XX", "X-", "-S", "-O", "-X", "- -".

new features to more traditional lexical and syntactic ones outperformed the baseline in classifying the simplified texts.

A further attempt to provide readability metrics with information about text cohesion has been favoured by the adoption of the Latent Semantic Analysis (LSA) model. LSA is a computational method for the representation of the content of a document as a vector in a multidimensional semantic space based on a large relevant corpus, which is adopted in a variety of applicative contexts, from information retrieval, educational technology and other pattern recognition issues where complex wholes can be treated as additive functions of component parts. Simplifying a lot (the interested reader is referred to Laundauer *et al*. 1998 and much related work), the rationale underlying this technique is that the level of semantic relatedness between texts or text subparts can be measured by looking at the specific contexts in which words tend to occur within the referent corpus: the higher the number of contexts that two words share, the higher their semantic relationship.

In a vector space model, the semantic similarity between two words or passages is then given by computing the cosine distance between the related vectors: the higher the cosine the more closely related the word or passage.

To make an example, the term *nurse* is likely to share many of the contexts of the word *doctor*, thus we can say that *nurse* and *doctor* have a strong semantic relationship. The advantage of LSA is that it also accounts for indirect relationships among words in contexts. This means that, if one looks at a text about hospitals where the word *doctor* exists but not the word *nurse*, the latter will be equally semantically related to the word *hospital* in light of the contexts that *nurse* shares with *doctor* and other words.

The general principles of LSA makes this model a valuable mean to approximate the inference load posed by a text: indeed, it permits to indirectly evaluate the amount of inferences a text requires, by looking beyond the simple number of words or arguments that overlap between adjacent sentences. On the assumption that higher cohesive texts display a higher semantic overlap among sentences or paragraphs, it is expected that these texts will report higher cosine values between all possible pairs of sentences. In this vein, LSA has been introduced among the variables of Coh-Metrix tool (§2.3.1.1), which indeed has been designed with a primary emphasis for the measure of coherence of writing texts.

2.3.3 Morpho-syntactic and syntactic features

While the distinctive attribute of traditional readability indexes was to exploit sentence length as a proxy of syntactic complexity (§ 2.2), the most recent automatic systems go far beyond this level by adding to their models a richer set of grammatical properties informed by linguistic research and sentence processing studies.

As for the other levels of linguistic description so far considered, the implementation of syntactic complexity features has made it possible by the use of more sophisticated parsers that perform shallow or deep analysis of text, depending on how well-formed the language structure of the target domain is expected to be. As described in Kate *et al.* (2010), such an approach enables to monitor fine-grained features qualifying the hierarchical structure of the syntactic tree, the most common of which are:

- Proportion of incomplete parses;
- Parse structure features;
- Average parse tree height;
- Average number of noun phrases per sentence;
- Average number of verb phrases per sentence;
- Average number of subordinate clauses per sentence;
- Syntactic similarity (i.e. the measurement of the uniformity and consistency of parallel syntactic constructions in text, typically at phrase level.)

The linguistic profiling of deep syntactic features deriving from the output of a morpho-syntactic and a syntactic parser represents a qualifying aspect of READ-IT (Dell'Orletta *et al.*, 2011), which is the first NLP readability assessment tool for Italian texts. Next paragraph is dedicated to illustrate more in detail the features on which READ-IT has been specialized in order to rate the degree of linguistic difficulty of a text and assign a readability score to it.

2.4 The READ-IT index

READ-IT[81] represents the first readability assessment tool based on NLP techniques for what concerns the Italian language.

[81] This software has been developed by the laboratory *Italian Natural Language Processing Lab* (ItaliaNLP Lab), which is part of the Institute of Computational Linguistics (ILC) "Antonio Zampolli" of the National

It has been designed with a view to a well-defined task, i.e. text simplification, and targets a specific readership, namely an audience with low literacy skills and/or mild cognitive impairment. A qualifying aspect of READ-IT is that it assigns a readability score both to the whole document and to each single sentence. The prediction of readability at sentence level is intendend to be a first step towards the process of text simplification, where it is essential to detect the areas of complexity within a text in a more fine-grained fashion (cf. Chapter 4).

The "core" of the READ-IT architecture is represented by the lexical, morpho-syntactic and syntactic features that the system is able to handle, thanks to the incorporation of an advanced suite of statistical NLP tools currently available for Italian: in particular, the PoS-tagger described in Dell'Orletta *et al.* (2009) and the dependency parser DeSR (Attardi, 2006), which are devoted respectively to the word category disambiguation and to the identification of syntactic dependency relations between tokens, using Support Vector Machine as learning algorithm[82].

Among the whole set of features which is possible to monitor from the automatic annotation, READ-IT has been trained against a subset of them which not only better digs into the notion of linguistic complexity, but also displays a greater connection with the intended audience, the language under examination and the reading object (document vs. sentence). Additionally, the selection of these features has been driven by computational issues, and specifically, their degree of reliable and large-scale identification by means of the available tools and resources.

The first category of features refers to the so-called "raw text" features, which stem from the tokenization process and resemble the variables typically used by traditional readability formulas, such as *Sentence Length*, calculated as the average number of words per sentence and *Word Length*, calculated in terms of characters per word.

The second class is composed by lexical features, i.e. features describing the composition of the vocabulary used in the text, whose detection relies on the output of the lemmatization and morpho-syntactic annotation. More specifically, this class accounts for the following parameters:

Council of Research (CNR) in Pisa. A demo of READ-IT is freely available at: http://italianlp.it/demo/read-it/ [last access: 01/07/2015]

[82] A more detailed description of the annotation pipeline underlying READ-IT will be provided in the next paragraph.

- the Type/Token Ratio (TTR): it measures the ratio between the number of lexical types (i.e. unique words) and the number of tokens, providing a metric of lexical variation[83]. Since it is known that such a measure tends to decrease as the text size increases, its calculation was limited here to the first 100 tokens and applied only to the level of the whole document, not the sentence.
- the Basic Vocabulary of Italian (BIV) rate features: this feature aims at capturing the psycholinguistic evidence that the more a text contains unfamiliar and specialist vocabulary (cf. 1.2), the harder is to comprehend. To measure the familiarity of the vocabulary, the authors relied on the *Basic Vocabulary of Italian* (GRADIT, De Mauro, 2000). This is a lexical resource that includes the about 7000 lemmas (both grammatical and content words), which are generally well-known to native Italian speakers; these words are further distinguished into three vocabulary ranges, covering the i) "fundamental words" (i.e. the first 2000 lemmas with top ranks in two frequency lists of Italian written (LIF) and spoken language (LIP)); the "high usage words" (i.e. the about 2700 subsequent words in these lists) and the "high availability words" (i.e. about 2000 relatively lower frequency words but still highly familiar to speakers as they refer to everyday objects and actions, e.g. fork, chair etc.);
- the Lexical Density: it measures the proportion of content-loaded words and it is calculated as the ratio between content words (nouns, verbs, adjectives and adverbs) over the total number of tokens in a text. Such a parameter has already been found to correlate with text complexity (Aluisio *et al.*, 2010) in "gold" corpora with predefined readability labels, and turned out to distinguish rich informative texts, whose cognitive demand is higher especially for low literacy readers. Psycholinguistic evidence, such the Gordon's studies on discourse processing we reviewed in paragraph 1.4.1, go in the same direction (at least, as far as the distinction between full lexical nouns and pronouns is concerned).

  However, it is worth remarking here the influence that the intended audience plays in the design of an automatic readability assessment model, as it might be the case that functional words, instead of content words, are particular demanding for impaired categories of readers (e.g. agrammatic aphasics).

[83] Beyond automatic readability assessment, it is worth remembering that TTR is also exploited in several domains of applied linguistics, e.g. child language research, where it is used as a benchmark to measure the richness/delay of vocabulary within a written or spoken production (see e.g Richards, 1987).

Moving further in detecting reliable proxies of complexity from deeper levels of linguistic annotation, the output of the PoS tagger allows capturing some of the indexes of morpho-syntactic complexity, which are likely to decrease text readability. In particular, READ-IT takes into account the following parameters:

- Language Model probability of part-of-speech (POS) unigrams: this feature is based on a unigram language model and it formalizes the assumption that the probability of a token is independent from its context; as shown again by previous works in the field (e.g. Pitler and Nenkova, 2008; Aluisio *et al.*, 2010), this is a reliable parameter to establish the degree of complexity from naturalistic corpora.
- Verbal features: this is a distinctive, language-specific feature of READ-IT, which was introduced to emphasize the "predictive" power deriving from the distribution of Italian verbs according to their mood, tense and person[84].

Finally, the algorithm checks those features which can better translate into a computational metric, based on a syntactic dependency formalism (see the next paragraph), some of the aspects of syntactic complexity emerged from the literature (§1.3). These features are the following ones:

- the unconditional probability of several types of syntactic relations (e.g. subject, direct object, modifier etc.);
- parse tree depth features, i.e. features extracted by quantifying three different measures:
    a) the depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency to some leaf;
    b) the average depth of embedded complement chains governed by a nominal head and including prepositional complements, nominal and adjectival modifiers;
    c) the probability distribution of embedded complement "chains" by depth.
- verbal predicate features, in particular:
    a) the number of verbal roots;

[84] Italian verbs have four finite (or verbal) moods (indicative, subjunctive, conditional and imperative) and three non-finite (or nominal) moods (infinite, participle and gerund). Within the finite moods, the Indicative has four absolute tenses: Present, Imperfect, Perfect and Simple Future. The Subjunctive has two absolute tenses, Present and Imperfect; the Conditional and Imperative have one absolute tense, Present. Each tense for each finite mood has six voices: first, second and third singular and first, second and third plural. For what concerns non-finite moods, the Participle has two tenses, Present and Past, while Infinite and Gerund have only the Present. These forms are invariable, except for the Participle that, under certain conditions, may take four forms marking gender and number.

b) the *arity* of verbal predicates, namely the number of dependents (both arguments and adjuncts) for each verbal head, that is given by looking at the number of dependency links sharing the same verbal head;
c) the distribution of verbal predicates by arity;
d) the percentage of verbal roots with elliptical subject, a structure compatible with the null-subject character of the Italian language.

- subordination features: subordination is typically associated with a higher degree of sentence complexity, as it increases the whole parse tree depth (cf. section 1.3). To account for the different impact of subordinate clauses on comprehension (deriving, among others, from the subordinate clause's position, internal structure and level of embedding), the following parameters have been assessed:
  a) the distribution of subordinate vs. main clauses;
  b) the order with respect to the main clause;
  c) the average depth of "chains" of embedded subordinate clauses;
  d) the probability distribution of embedded subordinate clauses "chains" by depth.

- length of dependency links: as discussed in paragraph 1.3.3, this is one of the most confident parameter of syntactic complexity deriving from sentence processing research. It is important to note that longer syntactic dependencies not only affect human readers' processing but negatively impact on the performances of statistical parsers, as demonstrated, among others, by McDonald and Nivre (2007). In READ-IT, this feature is calculated in terms of the total number of words occurring between the syntactic head and its dependent.

To implement these features, READ-IT approaches readability assessment as a classification problem. According to this paradigm, which represents the state of the art for this task, the algorithm infers a language model (i.e. a weighted representation of the distribution of the whole set of linguistic features) from the *gold corpus*; the latter, in this case, is a collection of texts previously labelled with the correct readability level to be dealt with. After the training stage, a new input document (i.e. text or sentence) will be assigned to a predefined readability category, on the basis of the similarity between its profile and that of the *gold corpus*.

In the current version, READ-IT assigns two readability categories, i.e. the "easy-to-read" and the "difficult-to-read", which have been modeled on two training corpora representative of two diverse varieties, both taken from journalistic genre, and distinguished in terms of text difficulty. The "pole" of simplicity has been defined on the periodic *Due Parole*[85] (Piemontese, 1996), since this a journal written in a controlled language and addressing a readership of people with low literacy skills or a limited cognitive impairment. The opposite pole is embodied by the national general-interest newspaper *La Repubblica*.

Finally, to provide the readability score, the system evaluates the degree of similarity between the profile of the new text (document or sentence) and either that of the easy- or the difficult-to-read corpus. This is done with respect to four different models, which have to be conceived as four different readability indexes of increasing complexity, resulting by a particular combination of the features described above. The easiest model is called **Base Model** and checks only raw text features; the second model is the **Lexical Model** and takes into account the distribution of raw and lexical features; the third model is referred to as **Syntactic Model** and it is based on a combination of morpho-syntactic and syntactic features and, finally, the **Global Model** combines a mixture of properties from all the previous models.

2.4.1 An overview of *LinguA* (Linguistic Annotation pipeline)

Any application rooted on natural language technologies for text analysis, such as automatic readability assessment, achieves the better performance to the extent it can rely on good linguistic annotations: briefly speaking, we can say that "good annotations support good applications" (Wilcock, 2009, p.1)[86].

Linguistic annotation is the process whereby the linguist information underlying a text is extracted and made explicit. This is generally a sequential process, in which the output of the prior level of annotation constitutes the input of the following one, in a way that tries to resemble how the human parser derives the meaning of a sentence[87]. The state of the art for this kind of task is represented by the use of supervised machine learning algorithms, which learn from a large hand-parsed training corpus the correct label to assign to a word, for each step of annotation.

[85] <http://www.dueparole.it> [last access: 01/07/2015]
[86] Graham Wilcock, *Introduction to Linguistic Annotation and Text Analytics.*
[87] At least in a modular approach to human language processing (cf. footnote: 4)

This is also the approach followed by *LinguA* (*Linguistic Annotation pipeline*), a suite of advanced statistical NLP modules[88], which represents the "backbone" of READ-IT.

LinguA incorporates three statistical modules: a tokenizer, a part-of-speech tagger and a dependency parser. They are arranged in a modular fashion, so that each component is fed by the input of the previous one and yields an output which is progressively more complex with respect to the level of linguistic information encoded (Montemagni, 2013).

The first component, TOKEN-IT, provides a first pre-processing of the text by dividing it into sentences (i.e. sentence splitting) and segmenting each sentence into orthographic units ('tokens'). The tokenized text is then enriched with morpho-syntactic information by adopting a part-of-speech (Pos) tagger (Dell'Orletta *et al.*, 2009), which is devoted to: a) attributing the grammatical label to each token (in according to a tagset comprising 14 coarse-grained Pos and 37 fine-grained Pos)[89] and b) assigning to each token the correspondent lemma. Finally, the DeSR parser (Attardi, 2009) runs the parsing of the sentence following a syntactic dependency approach. Within this paradigm (Nivre, 2007), the sentence is syntactically represented in terms of binary asymmetrical relationships, in which each token can play the role either of a head or a dependent.

Furthermore, the dependency arc linking the two tokens is labelled with the type of relation denoting the grammatical function that the dependent word has with regard to its governor[90].

Table 5 provides an example of the output of the whole annotation process.

Si fa presente che le mendaci dichiarazioni in atti pubblici e l'occupazione di immobili dichiarati inabitabili sono sanzionate penalmente. [91]
[Lit: It is pointed out that the mendacious declarations in public documents and the occupation of properties declared uninhabitable will be legally prosecuted.]

[88] *LinguA* has been jointly developed by the Institute of Computational Linguistics (ILC – CNR) and the University of Pisa. A demo version of it is available at: < http://linguistic-annotation-tool.italianlp.it/> [last access: 01/07/2015]
[89] For a complete description of the morpho-syntactic tagset, see Appendix I.
[90] For a complete description of the dependency tagset, see Appendix II.
[91] This sentence is taken from the corpus of bureaucratic texts described in Chapter 3, §3.3.

		Lemmatization	**Morpho-syntactic tagging**			**Syntactic parsing**	
Id	*Form (token)*	*Lemma*	*CPos*	*FPos*	*Morphological features*	*Syntactic Head*	*Type of depend.*
1	Si	si	P	PC	num=n|per=3|gen=n	2	Clit
2	Fa	fare	V	V	num=s|per=3| mod=i |ten=p	0	ROOT
3	presente	presente	A	A	num=s|gen=n	2	Pred
4	Che	che	C	CS	--	2	Arg
5	Le	il	R	RD	num=p|gen=f	7	Det
6	mendaci	mendace	A	A	num=p|gen=n	7	Mod
7	dichiarazioni	dichiarazione	S	S	num=p|gen=f	19	subj_pass
8	In	in	E	E	_	7	Comp
9	Atti	atto	S	S	num=p|gen=m	8	Prep
10	pubblici	pubblico	A	A	num=p|gen=m	9	Mod
11	E	e	C	CC	--	7	Con
12	l'	il	R	RD	num=s|gen=n	13	Det
13	occupazione	occupazione	S	S	num=s|gen=f	19	subj_pass
14	Di	di	E	E	--	13	Comp
15	immobile	immobile	S	S	num=p|gen=m	14	Prep
16	dichiarati	dichiarato	A	A	num=p|gen=m	15	Mod
17	inabitabili	inabitabile	A	A	num=p|gen=n	15	Mod
18	sono	essere	V	VA	num=p|per=3| mod=i|ten=p	19	Aux
19	sanzionate	sanzionare	V	V	num=p|mod=p|gen=f	4	Sub
20	penalmente	penalmente	B	B	--	19	Mod
21	.	.	F	FS	--	2	Punc

Table 5: Example of an annotated sentence in CoNNL format.

Table 5 is interpreted as follows. The second column ('form') reports the output of the tokenization process and is divided into as many rows as the number of word occurrences ('tokens') in the text; for each token, a progressive identifier number is assigned (first column, 'id'). Columns CPos and FPos specify the morpho-syntactic category of the word, along with some additional information which is associated with specific categories (e.g. number and gender features for common nouns): the latter is reported in column six ('*Morphological features'*). For instance, the token 'dichiarazioni' [declarations] (Id=7) has been associated with its lemma 'dichiarazione' [declaration], which is a common noun (S) and, more specifically, a plural (num=p) and feminine (gen=p) noun. The last two columns are dedicated to the syntactic annotation results: they identify, e.g., the noun 'dichiarazioni' (id=7) as the subject of the embedded passive verb 'sanzionate' [prosecuted] (id=19), which is the head of the dependency.

The same information provided in tabular format can also be graphically visualized, as shown in Figure 1: here the syntactic relationships are marked by labeled dependency arcs going from the head to the dependent.

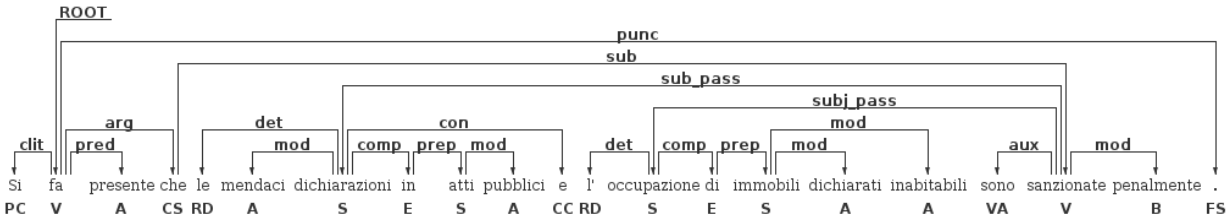


Figure 1: A graphical representation of the dependency syntactic annotation.

It is worth pointing out that the statistical tools underlying *Lingua* represent the state of the art for the Italian language, as testified by the results of the last campaigns devoted to the evaluation of the annotation tools for Italian (Evalita 2009, 2014)[92]. In particular, the morpho-syntactic tagger (Dell'Orletta, 2009) achieved a 96.34% accuracy[93] in simultaneously identifying the morpho-syntactic category and the associated morphological features. For what concerns the syntactic dependency analysis, the DeSR parser (Attardi *et al.*, 2009) reached 87.89% with respect to LAS[94] and 90.16% with respect to UAS[95] (Bosco *et al.*, 2014).

2.5 Towards genre-oriented readability metrics

While the development of general-purpose tools for automatic readability assessment has undergone a rapid growth within NLP-community, the evaluation of the impact of textual genre in readability evaluation is still at an early stage of investigation.

[92] http://www.evalita.it/2009/results ; http://clic.humnet.unipi.it/proceedings/Proceedings-EVALITA-2014.pdf [last access: 01/07/2015]
[93] The level of accuracy is calculated as the ratio between the number of tokens correctly classified over the total number of tokens analyzed.
[94] LAS (*Labeled Attachment Score*) accounts for the percentage of words that have been assigned the correct head and dependency label.
[95] UAS (*Unlabeled Attachment Score*) accounts the percentage of words that have been assigned the correct head.

A preliminary conclusion deriving from recent works in the field is that readability assessment is also "genre-dependent"; that is, the automatic evaluation of readability across different textual genres improves when genre-specific models are used to train the system.

With this respect, Sheehan *et al.* (2013) pointed out that traditional metrics like the Flesch-Kincaid Level score (§2.2) tend to overestimate the difficulty of literary texts and underestimate the difficulty of expository texts; based on this evidence, the authors developed distinct models for literary and expository texts, which were then used to train a readability assessment measure specialized for the characteristics of each genre. In a similar vein, Dell'Orletta *et al.* (2012) reported a higher degree of accuracy in assigning the correct readability level to texts belonging to different textual genres when dedicated models were used in training.

However, in order to train genre-specific readability models we need to collect appropriate resources for the language variety or textual genre to handle. This is especially true when such a variety may present lexical and syntactic patterns which are less, or not, instantiated in standard language, but yet required from a functional perspective.

Next chapter intends to investigate this issue by focusing on a specific textual genre, i.e. the bureaucratic language, which is known to be particularly challenging and hard to comprehend especially for non-expert readers. With the aim of distinguishing "genre-specific" markers of linguistic complexity from "unnecessary" stylistic features, which are typically instantiated in this typology of texts, a linguistic profiling investigation of a "quasi-parallel corpus" of Italian bureaucratic texts has been carried out by relying on a computational linguistics perspective to text difficulty analysis.

Chapter 3

Automatic readability assessment and the influence of textual *genre*: a corpus-based study focused on bureaucratic language

3.1 The bureaucratic language as a case-study: theoretical motivations

It is widely acknowledged that Italian bureaucratic language is a complex language variety. This is a fact that citizens can experience in their daily lives, when asked to fill in modules for school enrolment, obtaining residence or work permits, and so on.

Since the early nineties of last century, the problem of complexity in official writings has become a matter of interest also at institutional levels, under the accomplished awareness that a clear and plain communication is a primary mean to promote a real transparency and accessibility of public administration[96]. In this vein, Italian government, local administrations and academic research have been engaged in a "movement" towards simplification, which has given rise to a wide collection of recommendations, guidelines, style volumes, inspired to the *Plain Language* reforms deriving from Anglo-Saxon countries.

A common tenet underlying all these initiatives is the acknowledgment that the simplification does not have to result in a trivial and oversimplified text. This is because bureaucratic language contains some aspects of complexity compared to standard language, which cannot be always removed in any documents, because they depend on the complexity itself of the public administration, as well as on the performative nature of this language (Fioritto, 1997: 69; 2013: 151-158).

It is instead the use of "unnecessary" complex (whereas not completely inappropriate) lexical and morpho-syntactic choices (e.g. pseudo-technicisms[97], abuse of nominalizations, impersonal sentences) to be responsible of that obscure and far incomprehensible

[96] A strong impulse towards legal and administrative language simplification was also driven by the fundamental decision of the Italian Constitutional Court n° 364/88, which limited the effectiveness of the general principle that the 'ignorance of the law is no excuse' when the text law is unclear and possibly contradictory.

[97] For a definition of pseudo-technicisms, also called "collateral technicisms", cf. Serianni, (2005: 129).

communicative style, for which the famous Italian novelist Italo Calvino coined the term «antilingua»[98].

Starting from these considerations, the corpus-based study presented in this chapter has intended to investigate the peculiarities of bureaucratic language from an innovative perspective, based on the use of NLP-enabled features, with the aim of assessing whether, and to what extent, it is possible to characterize a model of linguistic complexity tailored to this genre, which would accomplish both theoretical awareness and computational treatability.

The chapter is structured as follows. Two introductory paragraphs anticipate the empirical part of this study: in §3.2 an overview of the features of bureaucratic language derived from qualitative research will be given, while §3.2.2 reviews the major steps towards the simplification of this language, which have been carried both within and outside government institutions. It follows the description of the corpus (§3.3) and the methodology adopted in this study (§3.4, 3.5), while section 3.6 is devoted to the corpus analysis results. In section 3.7, a qualitative inspection focused on relative clauses will be discussed.

3.2 The language of Italian public administration: linguistic peculiarities

An example of «sectorial but not specialized language» (Sobrero, 1993: 237), a kind of «special language»[99] (Cortelazzo, 1990: 5), a «complex variety merging the character of subcodes (or set of subcodes) to that of a formal register» (Berruto 1997: 164).

Any attempt to make Italian bureaucratic language understandable has to deal with the "hybrid" nature of this language variety[100], which is typically shared by the so-called *languages for special purposes* (LSP)[101].

[98] I. Calvino, *Per ora sommersi dall'antilingua*, Il Giorno, 8 february, 1965, than published again in *Una pietra sopra*. Discorsi di letteratura e società, Torino, Einaudi, 1980, pp. 122-126.

[99] By using the expression «lingua speciale», Cortelazzo (1994:8) defines «una varietà funzionale di una lingua naturale, dipendente da un settore di conoscenze o da una sfera di attività specialistici, utilizzata, nella sua interezza, da un gruppo di parlanti più ristretto della totalità dei parlanti la lingua di cui quella speciale è una varietà, per soddisfare i bisogni comunicativi (in primo luogo quelli referenziali) di quel settore specialistico; la lingua speciale è costituita a livello lessicale da una serie di corrispondenze aggiuntive rispetto a quelle generali e comuni della lingua e a quello morfosintattico da un insieme di selezioni, ricorrenti con regolarità, all'interno dell'inventario di forme disponibile nella lingua».

[100] Within the famous scheme proposed by Berruto (2012:17-24), the sociolinguistic variety of contemporary Italian can be characterized along four main dimensions, which account for changes over times (synchronic vs. diachronic variation), across place (diatopic variation), social groups (diastratic variation), medium of communication (diamesic) and communicative settings (diaphasic variation, sometimes referred to as "register" or "style"). The Italian bureaucratic language is positioned on the diaphasic variation axis and it is predominantly a written language variety.

As observed by Atkinson and Biber (1994: 356)[102], linguistic research on bureaucratic language is more frequently conducted for prescriptive purposes rather than descriptive ones, and research on bureaucratic Italian is not an exception. Such a perspective is certainly dependent on the simplification debate but seems also to derive from the awareness that the features really typifying the bureaucratic language as a special language *in a strict sense* (Berruto, 1997: 156) are considerably less than those of the special languages *in a broad sense* (Berruto, ivi).

As pointed out by Viale (2008:57), the features of the first type are mainly of lexical nature and account for a limited set of terminology domain, which is partially shared with legal language. The latter indeed is the most akin language variety, since public administrations not only are in charge of drafting normative texts, but they also guarantee their effective implementation (Fortis, 2005: 54-55)[103]: in light of this, it is especially the typology of administrative acts to rely on legal technicisms. Beyond this small intrinsic vocabulary, the complex activity of public administration makes it necessary to draw on the technical lexis of other special languages, in order to appropriately deal with a variety of matters, such as public health, education, environment, ICT technologies and so forth.

However, even if we limit to conceive bureaucratic language as a *special language in a broad sense*, we also need to acknowledge the presence of a large quantity of features, which are absent or underrepresented in standard language. It is crucial to note that such a kind of features do not seem to be constrained by any compulsory requirement of referentiality but are responsible, instead, of the "degeneration" of a bureaucratic text in *bureaucratese* (Lubello, 2014: 58), making this language artificial, obscure and far from being "reader-focused" (ivi, 113)[104]. An overview of these features is offered in table 1[105].

[101] It has to be noted that this term is not univocally accepted to define *special languages* in English sociolinguistics literature, too. Other labels, such as "language for specific purposes", "subcodes", "sublanguages" are also commonly attested.

[102] Quoted in Anderson J.,2006.

[103] Some scholars have pointed out that such a dependency is still so evident that bureaucratic language can be viewed as a «a sub-product of the language of the legislator», (P. Ungari, in Atti del Convegno "Il Linguaggio della divulgazione", II convegno, Milano, Selezione del Reader's Digest (1983: 55)).

[104] We do not tackle here the discussion about the motivations that still encourage the adoption of this language by public administrations, which are numerous and of different nature; some authors have explained it in terms of a misconception that everyday language in bureaucratic writings might degrade the role and "power" of public administration with respect to citizens (see Dardano, 1973: 185-188), but also more trivial reasons have been underlined, such as the incapacity of many public officials to "tune" their language to the actual receiver, who is rarely an expert, cf. Fortis (2005: 105-106).

[105] The table here illustrated is based on the most update surveys of Italian bureaucratic language provided by Viale (2008), Fortis (2005: 57-89), Lubello (2014: 45-61).

Lexical features	**Morpho-syntactic and syntactic features**	**Textuality features**
Pseudo-technicisms or *collateral* technicisms (e.g. *balneazione, fattispecie*)	Long and wordy sentences	Autoreferentiality (i.e. writer- vs. reader-focused writing)
Abstract nouns with –zione/-mento suffixes (e.g. *stipulazione, espletamento*), deverbal nouns, usually with zero suffix (e.g. *subentro, scorporo, utilizzo*) and denominal verbs (e.g. *relazionare, disdettare*)	Nominal style, deriving from: - large use of nominalizations instead of simple verbs; - verb periphrasis formed by a semantically empty verb + a deverbal noun (e.g. *apporre la firma*, *sottoporre a controllo* instead of the simple verbs *firmare*, *controllare*); - verb periphrasis formed by a semantically empty verb + lexical verb (e.g. *provvedere a riscuotere* instead of *riscuotere*); - abuse of indefinite moods in verb forms (infinitive, gerundive and, over all, past participle[106])	Fixed textual organization, typical of legal discourse
Archaic terms (e.g. *allorchè, testè, suddetto*) and latinisms (e.g. *una tantum, pro capite*)	Enclitic pronouns with finite verb	Extensive use of anaphoric and cataphoric links realized by means of specific adjectives, nouns and phrases (e.g. *visto, considerato, suddetto, sottoindicato, in calce*)
Forestierisms (e.g. *governance, back office, front office*)	Complex prepositional/conjunctive phrases (e.g. *ai fini di, dal momento che*)	Marked intertestuality (abuse of referents to external sources within a document)
Uncommon and formal terms (e.g. *diniego* instead of *rifiuto*)	Impersonal and passive sentences	Improper cohesion between sentences (e.g. when pieces of information are arranged in bullet or number lists, which separate the main clause from its dependent clauses, cf. (Fortis, 2005:77))

[106] Not surprisingly, participial clauses are among the primary distinguishing features of legal language at morpho-syntactic level.

Pleonastic and stereotyped phrases (e.g. *entro e non oltre, in riferimento all'oggetto*)	Parenthetic clauses and asides	Lack of coherence among sections and paragraphs.
Abbreviations and acronyms	Predominance of hypotaxis over parataxis, with long chains of subordinate clauses	
	Large use of negative sentences (e.g. *non saranno considerate le domande prevenute oltre* […] instead of the positive syntactic construction *saranno considerate solo le domande pervenute entro* […])	
	Non-canonical word order, e.g. with the adjective preceding the verb (e.g. l'*apposito modulo*), or less attested realizations of non-canonical orders (e.g. a sentence presenting a topicalized object without a resumptive pronoun in the main clause, e.g. *tali disposizioni riceveranno le amministrazioni*)	

Table 1: An overview of the features of *bureaucratese*.

3.2.1 Attempts towards the simplification of Italian bureaucratic language

As we mentioned in the opening paragraph, the impulse towards the simplification of administrative language – which can be viewed as an interdisciplinary topic among jurisprudence, linguistics and sociolinguistics (Fioritto, 2013: 149) – was part of a wider plan aimed at modernising Italian public administration (cf. Viale, 2005: 66-70), by also trying to reduce the gap between administrators and citizens. It was particularly between 1993 and 2002 that the most important initiatives were taken, under the impulse of the Italian Department of Public Service (Dipartimento della Funzione Pubblica).

The first real contribution towards bureaucratic language simplification was the *Codice di Stile delle comunicazioni scritte ad uso delle pubbliche amministrazioni*, a style volume published in 1993 by the then Minister of Public Function Sabino Cassese, which provides some basic recommendations for a proper use of language in administrative documents, by also taking into account the issue of gender in language usage. A separate section was also devoted to illustrating some possible rewritings of typical "bureaucratese" texts, a practice generally followed by all publications in the field.

Between 1994 and 1998, the aforementioned Department continued to take an interest in the language simplification problem, which was addressed by two dedicated projects carried

out by a research group composed by law experts, linguists and public officials, and coordinated by the jurist Alfredo Fioritto. The most remarkable outcome of this period was the publication of the *Manuale di Stile. Strumenti per semplificare il Linguaggio delle Amministrazioni Pubbliche* (1997), an enriched, more systematic and manageable version of the earlier *Codice* (cf. Tessuto, 2006: 414-421). The volume, specifically conceived to be used by ordinary readers rather than experts in language simplification, was composed by three parts: i) a guide for drafting administrative texts, which provided suggestions – at lexical, syntactic and textual level – for improving the clarity and readability of various typologies of administrative documents; ii) a glossary of 500 words typically used in bureaucratic language; iii) a guide for a correct layout of documents.

While both the *Codice* and, especially, the *Manuale* represented important governmental efforts towards the simplification of administrative language, it was only with the publication of the Directive of 2002 that the linguistic recommendations therein contained «acquire this time a more formal status»[107]. As noted by Fortis (2005: 96), by using the word "rules" instead of suggestions or recommendations, the Directive put a more prescriptive emphasis on the language question, which was additionally remarked by the statement that these rules had to be applied to all texts released by public administrations.

However, despite these promising contributions, the attention of governmental institution towards the question of language has undergone a dramatic downsize over more recent years. A discouraging signal of the new trend has been the recent publication (2014) of the new version of the *Codice di comportamento dei dipendenti pubblici delle pubbliche amministrazioni*, in which no recommendation for clear language is mentioned. On the other side, the loss of interest for such a crucial matter is counterbalanced by the more constant efforts conducted by academic research and interdisciplinary working groups. It is worth mentioning here the «Guida alla redazione degli atti amministrativi. Regole e suggerimenti», first published by the ITTIG-CNR Institute and Accademia della Crusca in 2011 (see footnote 2), which is the most update publication aiming to provide administrative personnel

[107] Dipartimento della Funzione Pubblica, *Direttiva sulla semplificazione del linguaggio dei testi amministrativi*, 8 May 2002. The spirit of this directive was partly anticipated by two previous measures: i) the «Codice di comportamento dei dipendenti delle pubbliche amministrazioni» (Decreto della Presidenza del Consiglio dei Ministri, Dipartimento della Funzione Pubblica, 10 april 2001), which states that «in drafting written texts and all other communications, the public officials shall adopt a clear and comprehensible language» [translation mine] and ii) the article 8, contained in the *Direttiva sulle attività di comunicazione delle pubblica amministrazione,* 7 February 2002 enacted by the same authority, which says more explicitly that «the communication of the public administrations shall accomplish the requirements of clarity, simplicity and conciseness, as well as guarantee the completeness and correctness of information» [translation mine].

with a accessible tool containing suggestions for enhancing the comprehension of their writings especially when devoted to layout people[108].

But what is the "content" of the simplification guidelines? In spite of the varying formulations, all the tools so far described contain similar suggestions to improve communication. A comprehensive summary of them has been conveyed in the form of a 30-point list ('trenta regole')[109], where the concept of linguistic clarity is explained by taking into account the multi-dimensional aspects of a text which might obstacle reading. According to the use of lexicon, e.g., the writer is recommended to:

- use everyday vocabulary, making reference to the *Basic Vocabulary of Italian* (BIV), (cf. paragraph 2.4) and avoid neologisms, latinisms, literary and archaic expressions, when not compulsory;
- limit the use of technical terms (i.e. words outside the BIV) and explain them whenever possible;
- limit abbreviations and acronyms, and ensure to give their extensive form when they are first mentioned;
- avoid ambiguous and polysemic words;

For what concerns **syntax**, the 'rules' suggest to:

- keep sentences short, if possible not beyond 20-25 words;
- prefer the use of active voice in verbs;
- connect words and sentences in a clear and short way, e.g. by also making explicit the subject and preferring finite rather than indefinite verbal moods;
- prefer positive, rather than negative, sentence formulations;
- prefer connecting propositions via coordination rather than subordination;
- avoid inserting parenthetical clauses and asides;
- prefer active rather than passive sentences;
- prefer concrete rather than abstract nouns and avoid nominalization;

[108] The Guide explicitly acknowledges the indications contained in previous academic manuals, among others, Franceschini/Gigli (2003); Cortelazzo/Pellegrino (2003); Raso (2005).
[109] http://www.maldura.unipd.it/buro/trentaregole.html [last access: 01/07/2015]

As it can be noted, the "easy-to-read" guidelines for bureaucratic writing comprise indications whose effectiveness goes beyond the domain they are primarily concerned with. They basically encourage the use of *Plain Language*, i.e. «the writing and setting out of essential information in a way that gives a co-operative, motivated person a good chance of understanding the document at first read, and in the same sense that the writer meant it to be understood» (Cutts, 1998)[110].

To a certain extent, such an inspiration to Plain Language – whose standards have been extensively modeled on the English language – is a potential drawback to take into account. As underlined by Viale (2008: 12-18) and Fortis (2004:51-61), not only these standards might be tricky to implement because of the intrinsic constraints posed by administrative documents, in terms of legitimacy and complexity, but a feedback from empirical studies testing the validity of these guidelines is still lacking for the Italian context. If some of these indications address phenomena that have a clear connection to the cognitive load underlying sentence comprehension, others should require adaptations to account for language-specific features, as well as for different categories of intended audience, e.g. native vs. L2 speakers. For instance, if we assume non local dependencies to be a marker of syntactic complexity (§ 1.3), the use of long parenthetical clause separating a head (e.g. a verb) from its arguments (e.g. the subject) is likely to add extra demands on working memory, thus making the sentence harder to parse. On the other side, the preference for passivization over the active voice might be favored in light of the information structure of a text passage, i.e. when the patient/theme is the topic of the sentence in the given context. This is likely to occur in bureaucratic documents, whose subjects are usually inanimate topics (such as regulations, law articles, bans, permissions, payments and so on) undergoing an action whose performer (e.g. an office, administration) is less prominent, if not irrelevant for the communication purpose.

The need of tailoring the guidelines to language-specific, as well as genre-specific, requirements is thus a relevant issue that should be better investigated by linguistic research.

[110] Quoted in: http://www.plainlanguagenetwork.org/resources/ [last access: 01/07/2015]

3.3 Corpus collection and description

The bureaucratic corpus on which the linguistic profiling investigation has been carried out is composed by 89 pairs of parallel texts, i.e. texts available both in their original and simplified version[111], (see Table 2). More specifically, the original texts are the authentic version of texts produced by several Italian public administrations, while the simplified texts consist in the rewriting of the original ones, which was manually performed by qualified linguists, as part of either academic courses on professional writing techniques or training courses specifically addressed to administrative personnel[112].

Although the corpus comprises diverse typologies of administrative texts, all the documents selected for the corpus were primarily conceived to be of major interest to people outside the administrative arena. For these texts indeed the need of balancing juridical adequacy on one side, and broader understandability on the other, poses significant, yet necessary, challenges to text readability assessment.

Nevertheless, assuming a perspective primarily oriented to the "external receivers" [113], i.e. layout people, is just a starting point in the attempt of classifying administrative textual typologies. Such a parameter identifies a macro-category of administrative acts, which contains in turn other sub-categories. The latter ones are classifiable along a continuum - which we can attribute to the «vertical dimension» of special languages (Cortelazzo, 1990) - that progressively increases the "juridical value" of the text towards a maximum of formality, which is a prerequisite of normative text. From the bottom to the top of this dimension, the typologies of administrative documents covered in our corpus are:

- set of forms;
- public announcements;
- advertising posters;
- reply letters;

[111] In what follows, we will often use the abbreviations *Bur_orig* and *Bur_simp* to indicate, respectively, the original and the simplified version of the corpus.

[112] The simplified versions here considered were mainly produced and made available by the Department of Linguistics at the University of Padua. Some of these texts (both the originals and the rewritings) belong to the so-called 'Corpus Tacs (Testi amministrativi chiari e semplici), which can be accessed for free through the link: <http://www.maldura.unipd.it/buro/index.html>, where the reader can also find extensive references on the topic of simplification and a survey of related works carried out by the aforementioned Department.

[113] It should be noticed that very few attempts of classifying textual typologies for the variety of bureaucratic language have been advanced: an interesting exception is represented by Viale (2008:103-109), who attributes an important role to the receiver as one of the variables along which discriminating administrative textual genres.

- letters to citizens (such as authorizations, concessions, orders);
- letters belonging to administrative proceedings;
- competition announcements;

A further discriminative variable within the corpus of original texts concerns the authority by which they were delivered, namely:

- Italian public Municipalities: this typology covers the majority of texts in the corpus, being also the most interesting one to monitor as it represents the straight communicative channel between local institutions and citizens.
- Ministry of Interior: it refers to one text only, which was published by the *Italian Ministry of Interior* to discipline political polls in 2006; more specifically, this text contains the institutional procedures that the members of electoral offices have to follow to carry out their job[114].
- Universities: the majority of these documents was produced as part of the project *Comunicazioni Istituzionali nelle Università. Raccolta di Modelli Testuali,* which was promoted by the Italian *Consorzio Interuniversitario sulla Formazione* (Co.Info.)[115]. The final purpose of this project was to provide Italian university administrations with a shared repertoire of standard models "written in plain, clear and effective language" to be used when communicating with the internal members (e.g. undergraduate, graduate and PhD students, teachers and other administrative staff).

It is worth pointing out that the collection of such a kind of corpus, which we propose to define as a "quasi-parallel monolingual corpus"[116], was particularly needed for the purposes of this investigation; indeed, it has provided an appropriate testing bed to monitor differences and similarities between two distinct language varieties – which we assume to instantiate two

[114] "Istruzioni per le operazioni degli uffici elettorali di sezione", Ministry of Interior, Department of Internal and Territorial Affairs, Rome, Istituto Poligrafico and Zecca dello Stato, 2006.

[115] For an overview of the project, the reader is referred to the website of the Consortium, at the following link: www.coinfo.net/documenti/RicercheIntervento/Allegati/Ricerca_Intervento_ComUniversità.pdf. [last access: 01/07/2015].
I wish to thank Prof. Michele Cortelazzo for making me available the electronic version of these texts, as well as those of the Italian public Municipalities taken from the book: Cortelazzo, M. A. (2005). *Il Comune scrive chiaro. Come semplificare la comunicazione al cittadino. Con 24 esempi di testi rielaborati e le istruzioni per scrivere con stile*, Santarcangelo di Romagna, Maggioli.

[116] This label aims at underlining that the process of simplification of the original texts, as it could be expected, turned out to tackle not only purely linguistic aspects but also high-level modifications concerning the structural organization, with the result that the corpus does not offer a perfect alignment at the level of sentence. As we will see in Chapter 4, the use of monolingual parallel corpora aligned at sentence level, is instead quite an essential requisite for current research in automatic text simplification.

different readability "poles" – specifically tailored to the bureaucratic language: the complex one is composed by the authentic texts released by public administration, while the simple one is embodied by the manually-created rewritings.

Bureaucratic Corpus	**Abbr.**	**N°of Documents**	**N°of Tokens**
Original texts	Bur_Orig	89	61.208
Simplified texts	Bur_Simp	89	43.780

Table 2: The «quasi-parallel» bureaucratic corpus.

3.4. Methodology

The bureaucratic corpus described in the previous paragraph has been investigated by adopting a linguistic profiling methodology, which consists in exploiting the output of different levels of automatic linguistic annotation (e.g. lemmatization, PoS tagging, syntactic parsing), with the aim of monitoring the occurrences either of single features, or combinations of features, in a text automatically pre-processed (van Halteren, 2004).

While the techniques to carry out linguistic profiling are today more sophisticated, thanks to the availability of computational linguistics tools and machine learning algorithms enabling a large-scale comparison between features, the general approach stems from classic corpus-based research on register variation (see, among others, Biber 1993, 1995, 1998; Conrad and Biber 2001) and follows the intuition that «linguistic features from all levels function together as underlying dimensions of variation, which each dimension defining a different set of linguistic relations among registers» (Biber, 1993).

As outlined in Montemagni (2013), a NLP-based linguistic profiling methodology requires two main ingredients: a wide range of linguistically motivated features, which can be searched for in the output of different levels of automatic linguistic annotation, and an appropriate reference corpus to compare to, in order to detect which features are most symptomatic for intercepting similarities and variations across different language varieties[117].

With respect to the first point, it should be noted that when linguistic profiling investigations are based on features derived from automatically parsed texts, the probability of error is a problem that researchers have to cope with. This is especially true when we investigate texts not belonging to the domain on which the statistical tools were trained or

[117] The term *language variety* has to be meant here in its broad sense, although it can be declined into several specificities along the dimensions described in Berruto (2012), cf. footnote 100.

developed on. In a seminal study, Gildea (2001) highlights that such a problem affects the performance of even state-of-the-art technologies, and especially statistical parsers, which undergo a considerable drop of accuracy when dealing with 'out-of-domain' texts. In light of these considerations, it is worth pointing out that the statistical tools for text analysis and features extraction, on which the present case study is based (§ 2.4.1), result from a previous specialization obtained by combining two training sets: the ISST-TANL Treebank[118], a dependency annotated corpus of 3,109 sentences taken from Italian newspaper articles and assumed as representative of standard language, and the TEMIS corpus (Venturi, 2012), a syntactically annotated corpus of Italian legislative and administrative texts. The latter should guarantee a more robust treatment of the linguistic peculiarities affecting bureaucratic language.

Besides, the methodology of linguistic profiling adopted for this study has already demonstrated to give promising results in corpus-based experiments aiming at monitoring different Italian language varieties (cf. Dell'Orletta *et al.*, 2013), thus corroborating the intuition that [*translation mine*] «the results of automatic linguistic annotation, despite being unavoidably subject to a margin of error, if appropriately explore, can offer reliable indications towards the reconstruction of the linguistic profile of a text» (Montemagni, 2013).

For what concerns the second "requirement" (i.e. the "monitor corpus" for comparison), we relied here on five corpora, each one belonging to a traditional textual genre, namely Journalism, Literature, Educational writing, Scientific prose and Legal language. Each of them has been distinguished into two sub-corpora, assumed as indicative of an easier and a more complex language variety. More specifically, for what concerns the first four corpora, the internal partition is conceived with regard to the potential reader that each sub-variety targets to. For arranging the legal corpus in a similar fashion, as we could obviously not rely on two legal language varieties specifically tailored to different categories of readers, we resorted to the Italian Constitution to embody the "easy" pole, in that prototypical of a simpler legal drafting[119].

Here follows a brief description of the corpora.

[118] The ISST-TANL is an updated version of the original Italian Syntactic–Semantic Treebank (ISST) (Montemagni *et al.*, 2003)).

[119] The legal corpus considered in this work is part of a broader corpus collected by Venturi (forthcoming), in which the author has investigated the linguistic profile of different typologies of Italian legal texts through a comparative and computational linguistics approach. Among the findings of her work, it has been proven that the Italian Constitution articles constitute an easier sub-variety of Italian legal language with respect to multiple NLP-enabled features.

The **journalistic genre** is represented by the same corpora which have been exploited for training READ-IT (cf. 2.4): the periodic *Due Parole* (Piemontese, 1996) for the simple variety, and the general-daily newspaper *La Repubblica*, for the standard Italian language. The **Literature genre** consists of two collections of novels, targeting respectively children and adults (here referred to as *Narr_child* and *Narr_adult*). The same holds for the **Educational genre**, represented by two different sub-corpora, including, respectively, materials used in primary (*Edu_child*) vs. high schools (*Edu_adult*). For the **scientific prose**, a collection of articles from Wikipedia (*Wiki*) has been compared to a corpus of scientific articles (*Scient_art*) on different topics (*e.g.* linguistics, climate changes etc.), in according to their potential reader, who is likely to be a non-expert vs. a specialist in the two cases. Finally, as mentioned before, the **legal corpus** gathers the Italian Constitution in its 1947 original version and a collection of legislative acts concerning environment issues. The corpora are detailed in table 3.

Genre	Corpus	Abbreviation	N° of documents	N° of tokens
Literature	Children Literature (Marconi et al., 1994)	Narr_child	101	19,370
	Adult Literature (Marinelli et al., 2003)	Narr_adult	327	471,421
			Tot: 428	**Tot: 490,791**
Journalism	La Repubblica (Marinelli et al., 2003)	La Rep	321	232,908
	Due Parole (Piemontese, 1996)	Due Par	322	73,314
			Tot: 643	**Tot: 306,222**
Educational materials	Educational Materials for Primary School (Dell'Orletta *et al.* , 2011b)	Edu_child	127	48,036
	Educational Materials for High School (Dell'Orletta *et al.*, 2011 b)	Edu_adult	70	48,103
			Tot: 197	**Tot: 96,139**
Scientific prose	Wikipedia articles from the Italian Portal "Ecology and Environment"	Wiki	293	205,071
	Scientific articles on different topics (e.g. climate changes and linguistics)	Scient_art	84	471,969
			Tot: 377	**Tot: 677,040**
Legal language	Normative acts on environment issues	Norm_acts		1,309,866
	Italian Constitution	It_const		10,487
				Tot: 1,320,353

Table 3: The reference corpora.

3.5 The selection of the features

As we said before, the linguistic profiling of the bureaucratic corpus here conducted has been stimulated by the attempt of contributing, from an original methodological viewpoint, to the identification of those patterns that define bureaucratic language as a "special language", since the status of this language with respect to standard Italian is still a matter of debate among scholars (cf. § 3.2). It also had a practical motivation in mind; indeed, we believe that it could ultimately serve as the starting point for the specialization of a general-purpose readability index like READ-IT (§2.4) towards the linguistic peculiarities of the bureaucratic prose.

More specifically, on the assumption that the two varieties of texts comprising the bureaucratic corpus here considered represent a "simple" and a "complex" model specific for this language variety, we wanted to investigate whether it was possible to detect both similarities and differences between them, by inspecting the multi-dimensional output of linguistic annotation. While the differences can be viewed as a computational metric to translate *bureaucratese*'s features – i.e. those stylistic constructs that have been affected in the rewriting process to enhance the readability of the original text (cf. §3.2) – the similarities are likely to be qualified as linguistic "markers" of this formal register, possible complex yet almost impossible to remove; this is especially true whereas they exhibit a sharp difference in comparison with the other reference corpora, and in particular the journalistic one which represent the variety of standard language.

With these concerns in mind, the features to monitor have been chosen so that to maximize:

- their explanatory "power" to formalize linguistic complexity predictors from a general-oriented perspective to automatic readability assessment research;
- their capability to make explicit *bureaucratese* markers as they have been highlighted by traditional, manually-carried, studies.

However, given also the data-driven approach of this exploratory study, in discussing the findings some remarks will be devoted to deepening certain tendencies empirically emerged, for which no prior examination has been devoted in the literature on bureaucratic language. Although some of them seem quite well instantiated (see *e.g.*, the different distribution of person features in verbal morphology), they will clearly need further investigation with a larger dataset.

An overview of the features considered for this study is given in Table 4. As it can be noted, they have been grouped according to the level of automatic linguistic annotation from which they derive, a distinction also adopted in their discussion. This approach allows us to discriminate the role, reliability and degree of sophistication that each level provides to linguistic profiling; however, it is only a descriptive methodology, as linguistic phenomena, especially when encompass syntactic properties, can be properly understood only by considering the interaction between low and high levels of analysis (consider, for instance, clausal subordination).

Feature typology	**Linguistic annotation level**	**Monitored features**
Raw text features	Sentence splitting/ Tokenization	Average sentence length
Lexical features	Lemmatization / Morpho-syntactic annotation	• Lexical density • Percentage frequency of unique types (lemmas) belonging to the *Basic Italian Vocabulary* (De Mauro, 2000) also with respect to the internal partition in the usage repertoires (fundamental, high usage, high availability) • Type/Token ratio
Morpho-syntactic features	Morpho-syntactic annotation	• (*Coarse* and *Fine-grained*) PoS distribution • Noun/Verb ratio • Morphological treats distribution
Syntactic features	Dependency parsing	Features based on the syntactic tree: • Average parse depth; • Average length of dependency links; • Average length of the longest dependency links; • Average length of prepositional chains;
		Features based on the syntactic dependency label: • Distribution of different dependency relationships
		Features concerning the use of subordination: • Main vs Subordinate distribution; • Average length of subordinate chains.

Table 4: The monitored linguistic features.

3.6 Linguistic profiling results: when the language of public administration turns into *bureaucratese*

We now illustrate the results of the linguistic profiling investigation, following the four-fold partition of features given in the previous paragraph. We will first give evidence of the similarities between the two varieties of bureaucratic texts (§3.6.1), thus discussing those patterns of features that turned out to distinguish the bureaucratic corpus "as a whole" and can be interpreted as qualifying a genre-specific notion of readability. In the second part of this section (§3.6.2), we will focus on the differences, i.e. *bureaucratese* "signatures" from a computational linguistics perspective.

3.6.1 Similarities between the bureaucratic language sub-corpora

The investigation into the peculiarities of the bureaucratic language moved from the output of the morpho-syntactic level of annotation. In particular, since «systematic differences in the relative use of core linguistic features provide the primary distinguishing characteristics among registers» (Biber, 1995: 136), we first looked at the distribution of the major morpho-syntactic categories across the corpora. This is visible in Table 5 (reported in Appendix III), which provides the percentage distribution of the 14 coarse-grained PoS covered by the underlying morpho-syntactic tagset[120].

3.6.1.1 The distribution of the "course-grained" morpho-syntactic features

At this level, a first indication that the language variety under analysis displays certain regularities as a whole comes from the data concerning the comparative distribution of the lexical (*i.e.* open-class word) categories: **nouns, verbs, adjectives** and **adverbs**. What we can observe is indeed the lowest percentage of adjectives and adverbs both in the bureaucratic corpus as a whole (respectively, Adj: 5.85 / Adv: 2.03), and in the two sub-corpora (*Bur_simp*, Adj: 5.72 / Adv: 1.91; *Bur_orig*, Adj: 5.98 / Adv: 2.14). Diametrically opposite is the distribution of nouns, which is one the most representative class across all the corpora, with similar percentage both in the overall corpus (30.17) and in the two sub-varieties

[120] We remember that the PoS tagger adopted here (Dell'Orletta, 2009) conforms to the ISST-TANL morpho-syntactic tagset, which distinguishes 14 coarse-grained PoS tags and 37 fine-grained PoS tags (see Appendix I).

(*Bur_simp*: 30.52 and *Bur_orig*: 29.82). In reference to the category of verbs, the bureaucratic genre shows a quite low frequency (average: 11.09; *Bur_simp* 11.05 / *Bur_orig*: 11.12), which is closer to the legal and the scientific corpora rather than to journalism, educational and literature, where the highest values are attested.

Shifting the attention to functional categories, bureaucratic language turns out to be distinguishable also with respect to the high frequency of **prepositions**. This finding was quite expected given the attested distribution of nouns: indeed, systematic associations between the occurrences of nouns, prepositions and attributive adjectives have been frequently reported by traditional multi-dimensional analyses focused on informative written registers (Biber, 1998). If our data confirm the correlation between nouns and prepositions, the same does not hold with respect to adjectives, which, as we noted before, are the less representative lexical category. Thus, in this case, a peculiarity of the bureaucratic genre is instantiated by a negative association pattern.

It is also worth noting that the high frequency of prepositions is expected to correlate in syntax with an elevated occurrence of prepositional complements modifying the noun head. However, to confirm this hypothesis, we clearly need to explore the output of the dependency parser (infra), where we should expect to find both complement and prepositional dependencies highly represented: according to the underlying dependency annotation, indeed the latter, identify the relation between a head (possibly a noun) and a prepositional complement while the former mark the relation between a prepositional head and its complement.

Figure 1 compares the distribution of nouns, verbs and prepositions among the corpora.

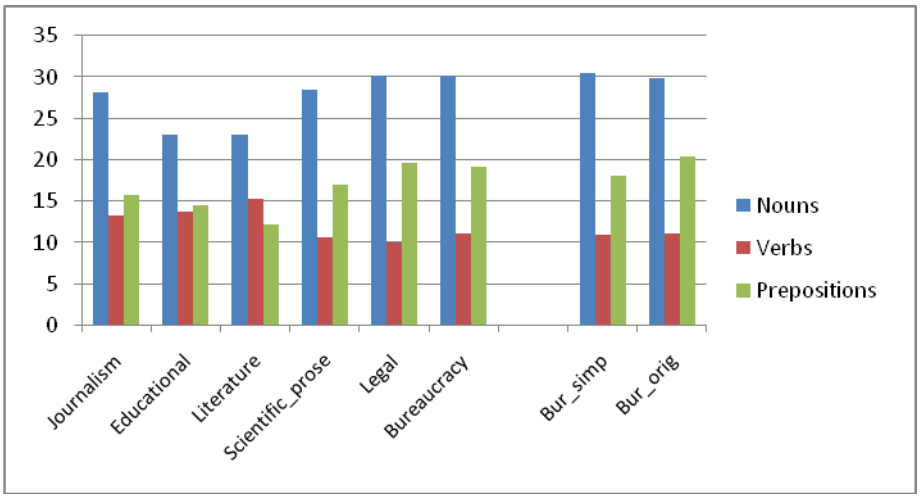


Figure 1: Percentage distribution of nouns, verbs and prepositions across genres (i.e. simple/complex varieties collapsed) and within the bureaucratic corpus (i.e. original and simplified sub-corpora).

3.6.1.2 Noun/Verb Ratio

Although the frequency of individual grammatical categories gives us some preliminary insights into language varieties, it is more interesting to consider the association between some selected parts-of-speech. With this respect, a measure which is highly informative in the literature on register variation is the **ratio between nouns and verbs**. Typically, a different pattern of nouns and verbs distribution discriminates genres and varieties along the diamesic dimension, with a higher percentage of verbs in speech productions and the opposite trend in written texts (Biber 1995, Voghera, 2005). This broad distinction has also proven to be effective in differentiating genres and registers within the two different diamesic varieties. For instance, focusing on written registers, fictional prose turns out to be closer to conversation as it is often structured in dialogic parts, while informative texts such as newspaper articles, tend to rely heavily on nominalizations, so they are the most distant to speech (Montemagni, 2013). Newswire texts have in turn a lower noun/verb ratio in comparison to academic prose and official documents (Biber, 1995).

Such a general tendency marking the correlation between the predominance of nouns and the marked informational focus of the text is also confirmed by our results (cf. figure 2): as expected indeed, the bureaucratic corpus reports high values of nouns/verbs ratio, very similar to those of scientific prose and only 0.35 percentage points less than legal language (3.07). Interestingly, no significant variation is detected between the percentage values of the simplified and the original sub-corpora, where instead *Bur_simp* (2.76) even exceeds by 0.08 points *Bur_orig* (2.68). These data confirm the highly-informative purpose of bureaucratic documents, the presence of numerous (and often mandatory) references to people, offices, institutions and so forth, which are conveyed by (common or proper) nouns, as well as the degree of abstractness that is typical of official writings, since they tend to be focused on inanimate topics (such as regulations, law articles, bans, permissions, payments and so on) rather than animate and specific subjects. (Raso, 2005: 112-113).

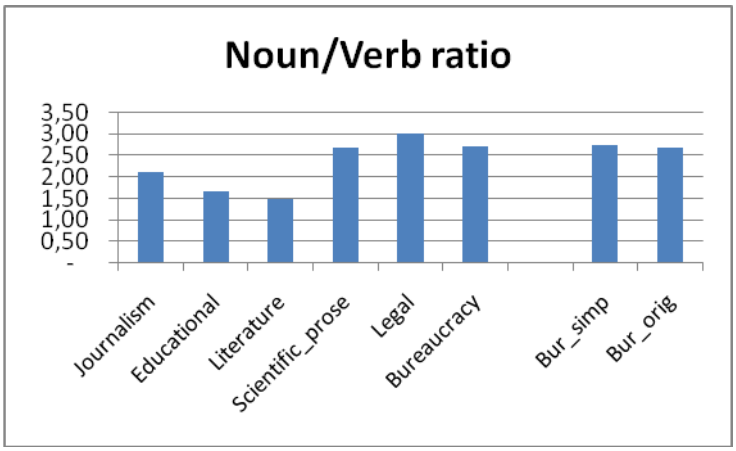


Figure 2: Noun/Verb ratio across major genres (i.e. simple/complex varieties collapsed) and within the bureaucratic corpus (*i.e.* original and simplified sub-corpora).

For each corpus, Table 6 reports the "coarse" values[121] of nouns/verbs ratio, i.e. the percentage between all the tokens which have been tagged, respectively, as nouns and verbs.

Genre	Corpus	N/V Ratio	
Journalism	Due Par	2.14:1	2.13:1
	La Rep	2.11:1	
Educational	Edu_child	1.55:1	1.67:1
	Edu_adult	1.81:1	
Literature	Lit_child	1.39:1	1.50:1
	Lit_adult	1.61:1	
Scientific_prose	Wiki	2.67:1	2.68:1
	Scient_art	2.69:1	
Legal language	It_const	2.62:1	3.07:1
	Norm_acts	3.52:1	
Bureaucracy	Bur_simp	2.76:1	**2.72:1**
	Bur_orig	2.68:1	

Table 6: Nouns/Verbs ratio percentage values across all the corpora.

Lexical features

The "predictive" power of lexical features for linguistic profiling was investigated by considering three parameters: i) the type/token ratio ii) the lexical density and iii) the degree of representativeness and internal distribution of the *Basic Italian Vocabulary*.

3.6.1.2 Type/Token ratio

[121] In according to the ISST-TANL morpho-syntactic tagset (Appendix I), the coarse-grained category of *nouns* is divided into three fine-grained tags: *abbreviations*, *proper nouns* and *common nouns*; the coarse-grained tag of *verbs* is further declined into *modal*, *auxiliary* and *main verbs*.

As reported in Table 7, the average value of *Type/Token ratio* in the bureaucratic corpus considered as a whole (0.69) is lower if compared to more descriptive genres, such as Narrative and Education; it is instead closer to the values retrieved in scientific prose (0.78) and in the legal corpus (0.46), which in turn is the nearest to the zero (*i.e.*, the value minimizing lexical variety).

Interestingly, when the attention goes to the internal distinction between *Bur_simp* and *Bur_orig*, the two values rather overlap (0.68 vs.0.70; t= -1.037962 p > .05).

This finding, again, goes in the direction of traditional qualitative analyses, which highlights the tight relationship that the language of bureaucracy holds with technical languages: not only it derives from legal language but also draws on the terminology that is proper of professional domains (such as health, construction industry, engineering, etc.) for dealing with a variety of different matters (cf. § 3.2).

As *special languages* are characterized by a high degree of formalism in designating their referent - a property known as *monoreferentiality* (Gotti, 2005) - it is highly difficult to convey the same meaning with synonyms or periphrasis, unless taking the risk of distorting the message or creating ambiguities. Therefore, if we assume type/token ratio to be a hint of lexical variety, the empirical data here collected make it possible to support the following claims: *i)* low type/token ratio values not only typify "easy-to-read" language varieties (consider, in this regard, the difference internal to the journalistic genre between *Due Parole* (0.66) and *La Repubblica*, (0.86)) but also reflect the precision required by technical discourse; *ii)* the original bureaucratic texts in our corpus were, on average, well-formed with respect to this parameter; iii) the type/token ratio value can offer a quantitative metric to verify the application a simplification rule formulated as follows: «an administrative act should avoid ambiguity and reach the greatest explicitness: hence it is suggested the use of the same term to designate the same action, concept or person, even at cost of high repetitions»[122].

Figure 3 provides a synthesis of the discussed results.

[122] *Guide for drafting administrative acts. Rules and suggestions*, p. 29.

Genre	Corpus	Type/Token Ratio (first 100 tokens)	
Journalism	Due Par	0.66	0.76
	La Rep	0.86	
Educational	Edu_child	0.80	0.80
	Edu_adult	0.81	
Literature	Lit_child	0.81	0.81
	Lit_adult	0.80	
Scientific prose	Wiki	0.77	0.78
	Scient_art	0.80	
Legal language	It_const	0.49	0.46
	Norm_acts	0.44	
Bureaucracy	Bur_simp	0.68	**0.69**
	Bur_orig	0.70	

Table 7: Comparison of the values of Type/Token Ratio (calculated for the first 100 tokens) across all the corpora.

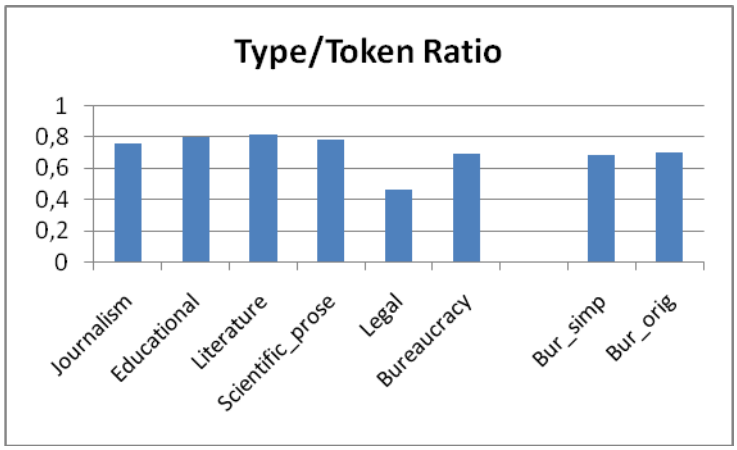


Figure 3: Type/Token ratio (calculated for the first 100 tokens) across major genres (i.e. simple/complex varieties) and within the bureaucratic corpus (*i.e.* original and simplified sub-corpora).

3.6.1.3 Lexical density

If type/token ratio gives an indication of lexical variety, lexical density works as a measure to quantify the informational load provided by a particular text, as it expresses the ratio between content and functional vocabulary. However, even limiting the attention to written registers[123], the explanatory power of lexical density in evaluating text complexity is not

[123] With regard to the diamesic dimension, a well attested finding is that spoken registers tend to be lexically lighter than written registers (Halliday, 1985). This is crucially related to the influence of extra-linguistic factors in oral communication: particularly, the time constraints in conversation prevent the speaker from planning utterances as accurately as he/she would do in writing, producing more frequently incomplete clauses, false starts, interjections etc.

straightforward, being it strongly related to the typology of texts, and eventually to the typology of the vocabulary.

Highly informative texts, such as the scientific and academic ones, tend to be lexically dense, as also confirmed by our data (see Table 8 and the related Fig. 4), showing that scientific prose exhibits the highest values of lexical density (0.577). Nevertheless, legal and bureaucratic texts are also connoted by a core informational load but they appear as the less lexically dense (Bur_orig: 0.538 Bur_Simp: 0.544; t=0.158, p>.05). Given the rationale behind this measure, which has also being used in automatic readability assessment research (cf. READ-IT features, § 2.4), another factor becomes crucially involved in characterizing the final rate: functional vocabulary. Functional words are indeed highly represented in our monitored corpus, especially with respect to the category of prepositions (cf. table 5, Appendix III), and they might be the source of low lexical density's values.

Genre	Corpus	Lexical Density	
Journalism	Due Par	0.564	0.564
	La Rep	0.564	
Educational	Edu_child	0.558	0.557
	Edu_adult	0.556	
Literature	Lit_child	0.568	0.573
	Lit_adult	0.578	
Scientific_prose	Wiki	0.584	0.577
	Scient_art	0.571	
Legal language	It_Const	0.555	0.543
	Norm_acts	0.533	
Bureaucracy	Bur_simp	0.544	**0.544**
	Bur_orig	0.538	

Table 8: Percentage values of lexical density across all the corpora.

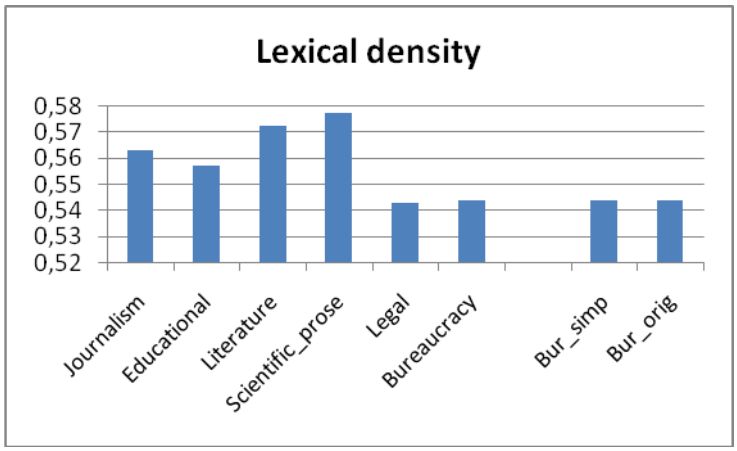


Figure 4: Lexical density across major genres (i.e. simple/complex varieties collapsed) and with respect to the internal distinction of the bureaucratic corpus (i.e. original and simplified sub-corpora).

3.6.1.4 Typology of vocabulary

A more qualitative indication of the degree of lexical complexity underlying bureaucratic texts comes from the results quantifying the *Basic Vocabulary of Italian* (BIV)'s rate (De Mauro, 2000) across the corpora. As Figure 5 shows, if legal texts are strongly characterized by the lowest percentage of lemmas belonging to the BIV (35.60), the bureaucratic language still continues to appear as a language for experts: the representativeness of the BIV is indeed much closer to that reported by the scientific texts and more than 10 percentage points less than the journalistic genre. Moreover, the similar height of the last two columns, depicting as usual the values of *Bur_simp* and *Bur_orig*, identifies this tendency as typical of the whole bureaucratic genre (*Bur_Sim*: 59.29; *Bur_orig*: 58.33; t=1.534 p>.05).

Genre	**Corpus**	**Percentage of words (in terms of different lemmas) belonging to the BIV**	
Journalism	Due Par	74.58	70.84
	La Rep	67.09	
Educational	Edu_child	74.57	73.57
	Edu_adult	72.56	
Literature	Lit_child	73.95	71.76
	Lit_adult	69.57	
Scientific_prose	Wiki	60.77	55.44
	Scient_art	50.11	
Legal language	It_Const	54.87	35.60
	Norm_acts	16.34	
Bureaucracy	Bur_simp	59.29	**58.81**
	Bur_orig	58.33	

Table 9. Comparison between the percentage of words belonging to the *Basic Italian Vocabulary* across all the corpora.

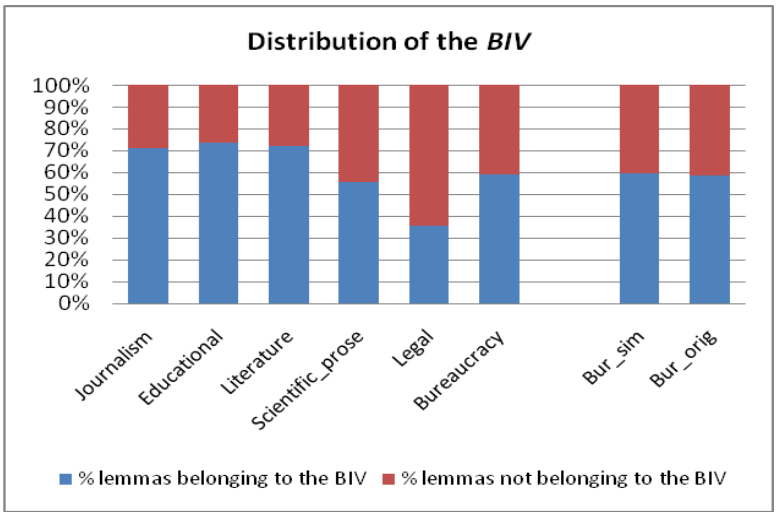


Figure 5: *Basic Italian Vocabulary* across major genres (i.e. simple/complex sub-varieties collapsed) and with respect to the internal distinction of the bureaucratic corpus (i.e. original and simplified sub-corpora).

Further interesting details come from the data reporting the internal distribution of the BIV, which, we remind here, is divided into three usage classes[124]. In particular, Figure 6 suggests similar conclusions with respect to the data of the "fundamental vocabulary", although with a slight internal difference that is worth pointing out: as highlighted by the circle, indeed, the percentage of words belonging to the fundamental vocabulary is a little higher in the simplified register (67.12%) than in the original one (Bur_simp: 67.12%; Bur_orig: 64.94%; $t=2.986$ $p<.01$), a tendency that, we can speculate, indirectly witnesses the efforts made by the writers of the simplified texts to rely, whenever possible, on more familiar words in according to the easy-to-read guidelines[125].

Genre	Corpus	Percentage of words belonging to the Fundamental Vocabulary	
Journalism	Due Par	75.06	73.53
	La Rep	72.00	
Educational	Edu_child	73.02	73.15
	Edu_adult	73.29	
Literature	Lit_child	76.84	76.31
	Lit_adult	75.78	
Scientific_prose	Wiki	68.18	67.04
	Scient_art	65.89	
Legal language	It_Const	70.03	66.69
	Norm_acts	63.36	
Bureaucracy	Bur_simp	67.12	**65.63**
	Bur_orig	64.94	

Table 10: Comparison of the percentage of words belonging to the *Fundamental Vocabulary* across all corpora.

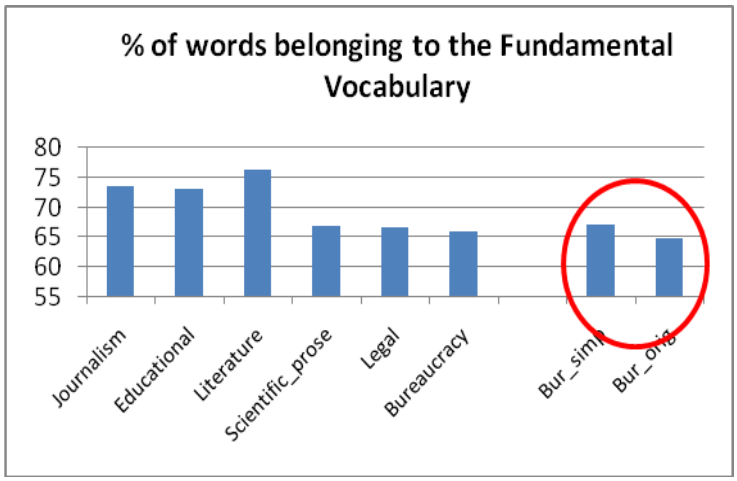


Figure 6: *Fundamental Vocabulary* across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

[124] Fundamental, high usage and high availability vocabulary; cf. par. 2.4.
[125] «Therefore, whenever possible, words from the basic vocabulary should be chosen, preferring those to more rare words.», *Guide for drafting administrative acts*, cit, p. 25.

Syntactic Features

The next three paragraphs will enrich the reconstruction of the linguistic profile of the bureaucratic corpus by focusing on the output of the dependency parser, which formalizes the syntactic structure underlying a text in terms of binary asymmetrical relationships between tokens, namely a head and its dependent(s) (cf. § 2.4.1). The whole ouput of this level is reported in Appendix IV (Table 11).

Although the precision of the automatic linguistic analysis is less accurate at this level, especially when dealing with texts outside the domain of training data (cf. § 3.4), syntactic structure is surely the most informative domain to characterize text difficulty. Indeed, it allows exploring the role of those factors (such as the length and depth of the syntactic dependencies), whose impact on text readability has also a stronger cognitive relevance (cf. section 2.3). Consequently, the output yielded at this level, if properly investigated, enables us to reconstruct the syntactic profile of a text, not only by quantifying the distribution of different typologies of dependencies, but also by making explicit their internal structure.

With this respect, dealing with a *quasi-parallel* corpus has been particularly fruitful as it has proven the existence of some syntactic tendencies that typify the bureaucratic prose as a distinct variety, without being affected by the simplification process. We refer, in particular, to the use of prepositional complements, the structure of nominal modifiers and the preference for a hypotactic prose, which will be discussed in the next three paragraphs.

These similarities were actually the less predictable as it has been widely emphasized that *bureaucratese* tends to be characterized by very infrequent syntactic patterns, which are responsible of a poor clarity; thus, we expected that they would have been removed after the simplification. From this perspective, uncovering these common features, some of which are crucially related to a general (*i.e.* genre-independent) notion of syntactic complexity, becomes crucial for training a readability model specialized to the peculiarities of the bureaucratic syntax: they seem indeed to suggest the need of tailoring, if not the parameters themselves, their minimum and maximum potential values within this textual typology.

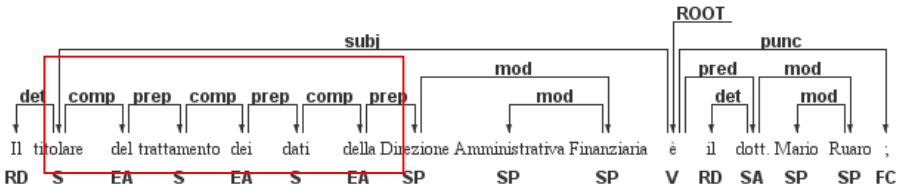
3.6.1.5 Prepositional complements

A first step towards the characterization of the syntactic profile of the bureaucratic corpus was the comparison between the most instantiated dependencies in *Bur_Simp* and *Bur_orig*. In this regard, it is interesting to observe the similar percentage distribution obtained by two

kinds of dependencies, *i.e.* "comp" and "prep" (cf. Table 11, Appendix IV), whose higher occurrence directly correlates with the findings of the PoS unigrams (cf. Table 5, Appendix III).

At that level of analysis, it was shown that prepositions were a distinctive feature of the whole bureaucratic corpus, predicting a likewise **high occurrence of prepositional complements** in syntax. In accordance to the underlying annotation tagset, prepositional tokens represent in fact the head of a "prep" relation and the dependent of a "comp" relation, as graphically illustrated in (1).

(1) Il **titolare del trattamento dei dati della Direzione Amministrativa Finanziaria** è il dott. Mario Ruaro.
[The *data controller of Financial Administrative Office* is Dr. Mario Ruaro]



The high frequency of prepositional complements depends, in turn, upon the "core" nominal style affecting bureaucratic (and, more in general, highly informative) writings[126], as they represent the typical syntactic realization of nominal modification.

In order to enhance the investigation into the structure of nominal modifiers we thus calculated the **average length of prepositional chains**. With this respect, it is important to clarify that in a syntactic dependency formalism, a prepositional chain has to be intended as the counterpart of what a complex noun phrase is in a phrase-structure representation: it describes a sub-tree that is instantiated by a sequence of tokens where a nominal head governs a single (or more hierarchically-embedded) prepositional modifiers(s). The attention for these structures directly relates to syntactic complexity, as they capture the frequency of heavy noun phrases, which represent an established source of sentence processing difficulty.

The results detailed in Table 12 show that the bureaucratic corpus contains very long prepositional chains (1.55), only 0.07 percent points less than the score attested in legal corpus (Figure 7).

[126] Cf. the values of nouns/verbs ratio and the percentage distribution of nouns reported in table 3.

Interestingly, this tendency was not particularly affected by the attempts of rewriting the original texts into a simplified version (Bur_orig: Bur_simp; t= -1.037, p>.05), suggesting that 'longer-than-the-average' prepositional chains (cf. the mean score obtained by the journalistic corpus) represent the correspondence, at syntactic level, of a prose largely arranged around nouns. These data seem to reflect another genre-specific pattern of complexity for the domain of texts under examination.

Genre	**Corpus**	**Average length of embedded prep. chains**	
Journalism	Due Par	1.24	1.29
	La Rep	1.35	
Educational	Edu_child	1.17	1.21
	Edu_adult	1.25	
Literature	Lit_child	1.17	1.17
	Lit_adult	1.17	
Scientific_prose	Wiki	1.36	1.40
	Scient_art	1.45	
Legal Language	It_Const	1.37	1.60
	Norm_acts	1.83	
Bureaucracy	Bur_simp	1.51	**1.53**
	Bur_orig	1.56	

Table 12: Percentage values of the average length of embedded prepositional chains across all the corpora.

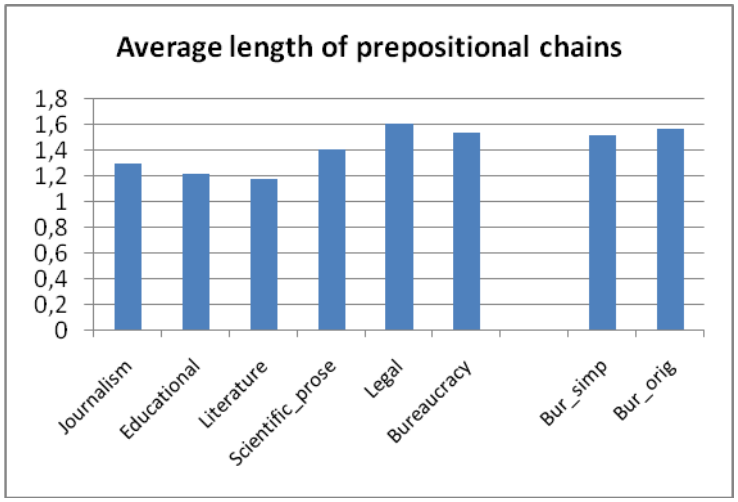


Figure 7: Average length of embedded prepositional chains across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

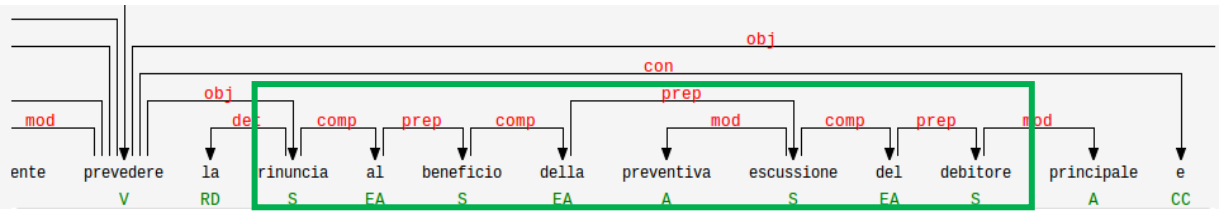
However, two caveats are important for a better comprehension of the predictive power of this feature in assessing syntactic complexity in texts. First, there is an additional piece of information that is worth comparing in our corpora: the probability distribution of embedded

prepositional "chains" by depth. The syntactic dependency output, indeed, not only returns a numerical score of the average length of prepositional chains, but also distinguishes them according to their hierarchical level of embedding: specifically, it enables for the extraction of the percentage distribution of sequences made of 1, 2, 3, or more complements and dependent on a noun. This is an important feature, as deeper embedded complement chains overload sentence processing and impact on the difficulty of a text (cf. § 1.3.2). As we will discussed in section 3.6.2, if the two sub-corpora are comparable with respect to the average length of prepositional chains, the same does not hold for the distribution of these structures according to their level of embedding.

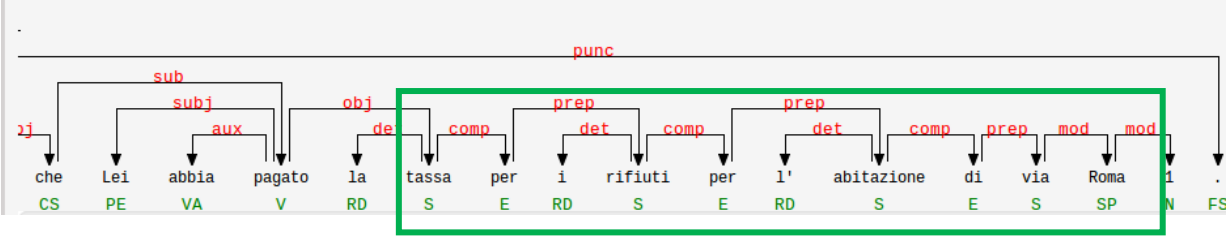
Second, it is worth remembering that the syntactic representation of a prepositional chain as that described in example (1) captures the dependency relationship between a nominal token and a (sequence of) complement modifier(s), but does not distinguish the lexico-semantic features of the noun head. The neutralization of the noun's argument structure is a potential shortcoming for a study aimed at training an advanced genre-specific readability index. For instance, two same-length prepositional chains, governed respectively by a nominalization or a deverbal noun (example 2) and a simple noun (example 3), might have a different impact on the informational load conveyed by the sentence. Only the former in fact may include the same types of arguments/adjuncts as the corresponding verbal entry, thus requiring more processing resources in order to reconstruct the event-based structure.

(2) "Si precisa che, a norma dell'art. 30 della legge n. 109/1994 e successive modifiche e integrazioni, la polizza fideiussoria e la fideiussione bancaria dovranno espressamente prevedere **la rinuncia al beneficio della preventiva escussione del debitore principale** […].

[Lit: It is specified that, as enacted by PL 109/1994 c. 30 and subsequent modifications and integrations, the insurance policy and the bank guarantee must explicitly ensure *the waiver of the benefit of the preventive enforcement of the principal debtor* […]"



(3) Non ci risulta che Lei abbia pagato la **tassa per i rifiuti per l'abitazione di via Roma 1**."
[Lit: We have no reports that you have paid *the tax for waste for the household in 1, Street Roma*.]



3.6.1.6 The subordination

A separate analysis was dedicated to monitoring the use of subordination within the corpora. This is certainly one of the most agreed-upon factors of text complexity, although with the limits that a broad notion of subordination is exposed to. We know that subordination increases the amount of the embedding to be computed and the propositional content of the whole sentence (§ 1.3), however, [*translation mine*] «not all the subordination is alike: it is not the presence of a subordinate clause in itself that constitutes a strong element of complexity but the combination between subordination and other factors, such as the relative order between the main and the subordinate clause; the degree of embedding of the subordinate clause; the correspondence between the events chain and the sequence of clauses» (Voghera, 2001:69)).

Some of these constraints can be reliably inspected from a dependency parsed text. With this respect, we have first calculated the **ratio between subordinate and main clauses**, on the assumption that the higher this value, the richer the informative content of the sentence.

According to the adopted dependency tagset, the main clause is uniquely identified by a verbal root, while to quantify the degree of subordination we need to consider the presence of:

i) non-subject adverbial clauses, *i.e.* clauses subcategorized by the main verb, and thus linked to it through an "arg" dependency, which can be introduced by either a preposition (infinitive subordinate clauses) or a conjunction (finite subordinate clauses);

ii) clausal modifiers, *i.e.* clauses which modify the main clause by providing a more precise semantic value, and thus linked to the verbal root by the general "mod"

dependency and some more specific clause modifiers (es. locative modifier, "mod loc") or temporal modifier, "mod_temp".

The comparison of this parameter among the corpora proved that the language of bureaucracy makes extensively use of hypotactic constructions (Figure 8), as demonstrated by the highest ratio between subordinate and main clauses, which is recorded both in the average corpus (0.58) and within the two sub-corpora (*Bur_simp*: 0.56; *Bur_orig*: 0.60 t: -0.170407 p>.05) (Table 13).

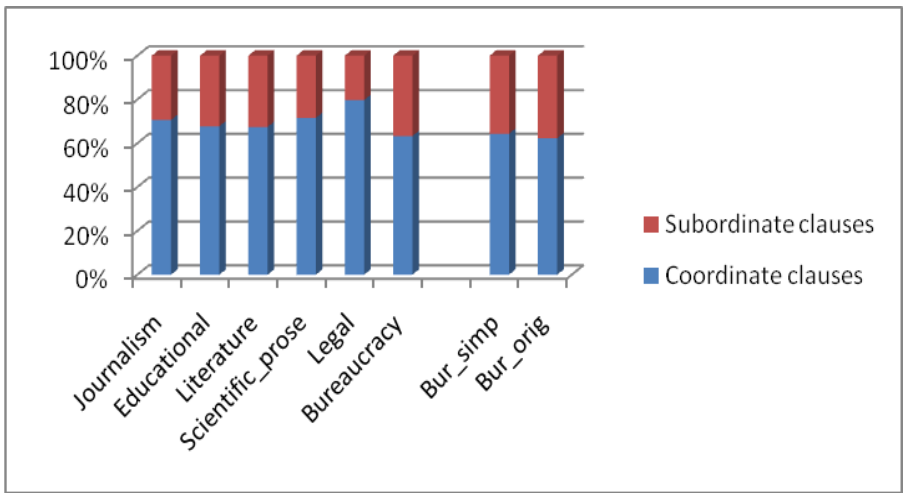


Figure 8: Main/Subordinate clauses ratio across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	Main Clauses	Subordinate clauses	Sub/Main ratio	
Journalism	Due Par	73.55	26.14	0.36	0.42
	La Rep	67.33	32.36	0.48	
Educational	Edu_child	69.94	28.42	0.41	0.48
	Edu_adult	63.55	35.04	0.55	
Literature	Lit_child	68.32	30.69	0.45	0.48
	Lit_adult	65.77	33.92	0.52	
Scientific_prose	Wiki	72.92	26.74	0.37	0.40
	Scient_art	69.12	29.70	0.43	
Legal language	It_Const	86.07	13.93	0.16	0.26
	Norm_act	73.39	26.61	0.36	
Bureaucracy	Bur_simp	63.63	35.24	0.56	**0.58**
	Bur_orig	61.58	37.29	0.60	

Table 13: The percentage distribution of main and subordinate clauses in all the corpora.

In addition, we can observe that the **average length of subordinate clauses chains**[127] within the two sub-corpora is almost the same (Table 14 and Figure 9): 0.96 in *Bur_orig* and 0.95 in *Bur_simp*.

Genre	Corpus	Average length of subordinate clauses chains	
Journalism	Due Par	1.01	1.09
	La Rep	1.17	
Educational	Edu_child	0.98	1.08
	Edu_adult	1.17	
Literature	Lit_child	1.20	1.16
	Lit_adult	1.11	
Scientific_prose	Wiki	0.90	1.03
	Scient_art	1.17	
Legal Language	It_Const	1.03	1.11
	Norm_acts	1.18	
Bureaucracy	Bur_orig	0.95	**0.95**
	Bur_simp	0.96	

Table 14: Percentage values of the average length of subordinate clauses chains across all the corpora.

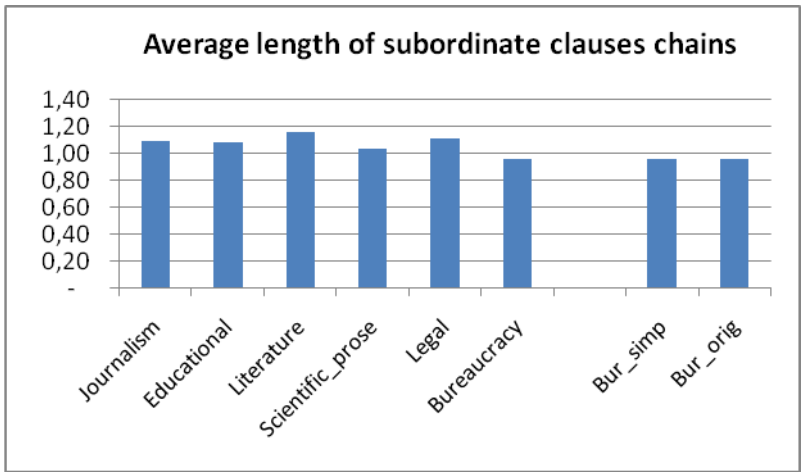


Figure 9: Average length of subordinate clauses chains across genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

If these data confirm that hypotaxis is a typical feature of bureaucratic prose, when we compared the two sub-corpora with respect to finer parameter digging into the degree of hierarchical embedding of the syntactic dependencies and the typologies of subordinate

[127] Given a dependency parsed text, these data can be made available by focusing on the internal structure of the sub-tree identifying a subordinate dependency and, more specifically, by considering the number of recursively embedded subordinate nodes which depend on the first parsed subordinate head, i.e. the head directly governed by the matrix verb. It is worth underlining that such data are here underestimated, as they do not account, e.g, for relative clauses (see § 3.7).

clauses, some significant distinctions emerged: we will account for these in the next paragraph, where the attention will be specifically drawn to the characterization of *bureaucratese* syntactic markers.

3.6.2 What about *bureaucratese*?

The same approach adopted so far has enabled us to investigate the second question underlying this study, which can be formulated as follows: is it possible to rely on the linguistic profiling methodology as a mean to infer the manual simplification interventions carried out on the original text with respect to the language variety under examination?

An answer to this question should require the identification of the main linguistic patterns along which the two profiles diverge, with a special interest to those patterns that better seem to provide concrete evidence to the qualitative indications towards a plain and clear administrative writing (§3.2.1).

As in the previous section, the features of interest will be checked and described against the output of different levels of linguistic description.

3.6.2.1 Average sentence length

As it was fairly expected, a first indication of a substantial misalignment between the two bureaucratic language sub-varieties is provided by a rough parameter of sentence complexity, i.e. the **average sentence length** (see Figure 10).

If we restrict the attention to the whole bureaucratic corpus compared to the others, this feature does not seem to play a salient role as a genre marker. Crucially, when the internal distinction is taken into account, a different scenario is offered and *Bur_orig* shows to contain considerably longer sentences than the simplified one (Bur_orig: 26.72; Bur_simp: 20.00; $t=6.046991$, $p< 0.01$).

Despite the direct relation between the linguistic complexity of a text and the number of words it consists of[128] (a correlation capitalized by traditional readability formulae), the difference here observed within the same register is a clear symptom of that tendency to

[128] The *Mean Length of Utterance* (MLU) is a rough measure of sentence complexity exploited within different fields of applied linguistics; for instance, from the perspective of language acquisition research, a MLU (typically measured in morphemes) below a given threshold is used as a marker of language impairment or delay.

prolixity which is typically reported among the hints of *bureaucratese* and, consequently, highlighted by the plain writing guidelines as an aspect to pay attention to.

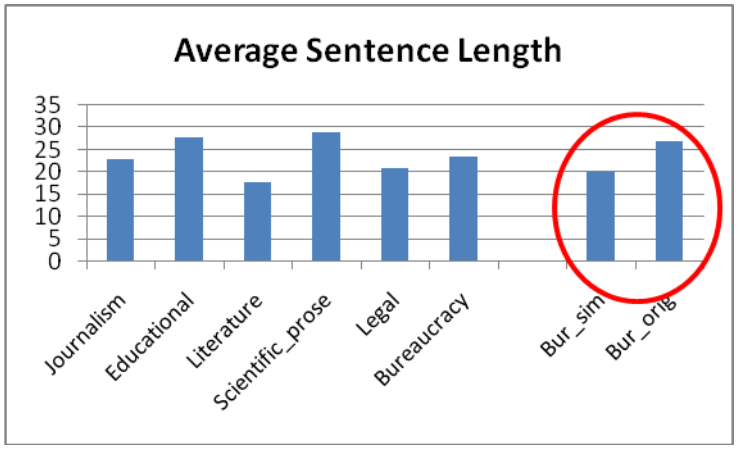


Figure 10: Average sentence length across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	Average Sentence Length	
Journalism	Due Par	19.20	22.87
	La Rep	26.54	
Educational	Edu_child	23.64	27.64
	Edu_adult	31.63	
Literature	Lit_child	16.96	17.61
	Lit_adult	18.25	
Scientific_prose	Wiki	25.80	28.73
	Scient_art	31.65	
Legal language	It_Const	16.59	20.79
	Norm_acts	24.99	
Bureaucracy	Bur_simp	20.00	23.36
	Bur_orig	26.72	

Table 15: Percentage values of the average sentence length across all the corpora.

3.6.2.2 The distribution of the "fine-grained" morpho-syntactic features

By looking at the frequency of the major grammatical classes (*i.e.* nouns, verbs, prepositions) (Table 5, Appendix III), it was possible to discover the presence of regularities between *Bur_orig* and *Bur_simp*. However, the fine-grained level of the morpho-syntactic annotation also reveals some interesting patterns of variation, which can help us refining the notion of linguistic complexity for this particular domain.

Let's consider, e.g., the distribution concerning the category of conjunctions. Although at the broad level, the percentage distribution of conjunctions has not appeared particularly

relevant in characterizing the overall bureaucratic corpus, *Bur_Simp* and *Bur_orig* slightly diverge with respect to the proportion between **coordinating and subordinating conjunctions** (*i.e.* the two subcategories provided by the adopted tagset for labeling conjunctions), with a prevalence of subordinating conjunctions in *Bur_Simp* (0.99) than in *Bur_orig* (0.77), (t= 1.99 p< .05).

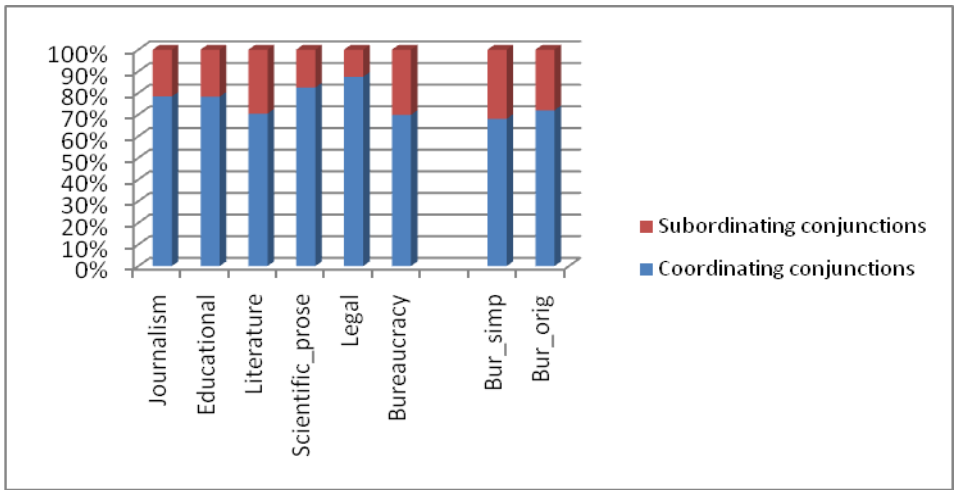


Figure 11: The proportion between coordinating and subordinating conjunctions across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	Sub. Conj.	Coord. Conj.	Sub./Coord. Ratio	
Journalism	Due Par	0.67	3.02	0.22	0.79
	La Rep	0.91	2.70	0.34	
Educational	Edu_child	0.99	3.39	0.29	1.02
	Edu_adult	1.06	3.96	0.27	
Literature	Lit_child	1.41	3.43	0.41	1.36
	Lit_adult	1.31	3.05	0.43	
Scientific_prose	Wiki	0.57	3.18	0.18	0.63
	Scient_art	0.69	2.75	0.25	
Legal language	It_Const	0.72	4.57	0.16	0.59
	Adm_acts	0.45	3.68	0.12	
Bureaucracy	Bur_simp	0.99	2.10	0.47	0.88
	Bur_orig	0.77	1.98	0.40	

Table 16: Proportion between coordinating and subordinating clauses across all the corpora

We might look at these data as a preliminary indication in the attempt of shedding light on the different ways in which subordination is expressed within the two varieties. A more consistent use of subordinating conjunctions seems to suggest that the authors of the simplified texts have relied, to a more extent, on explicit subordinate "markers" (e.g. that, while, as, if etc.) to hierarchically arrange complements around the main clauses (see, for

instance, examples in (4a, 4b, 4c)). In the original corpus, instead, subordination was more frequently introduced either by complex prepositional locutions (cf. (5a, 5b)) - which are typically conveyed in Italian as a multi-word expression composed by the sequence preposition/common noun/preposition, e.g. *ai fini di/allo scopo di* (= in order to), *in riferimento a* (= with reference to)) - or by implicit clauses modifying the main clause (cf. (5c)).

(4)

a) ***Se**, invece, preferite mantenere il regime del diritto di superficie*, vi ricordiamo che la convenzione preliminare che avete sottoscritto vi obbliga a chiedere al Comune di Schio l'autorizzazione preventiva per qualunque passaggio di proprietà, affitto, cambio societario, ecc.

[Lit: *If, instead, you prefer to hold the leasehold regime*, we remind you that according to the preliminary convention which you have signed you need to require prior authorization to the Municipality of Schio for any transfer of property, rental, corporate change, etc.]

b) I proprietari di autoveicoli e i titolari di patente non sono obbligati a cambiare l'indirizzo su libretto di circolazione e patente, ***perché** l'obbligo è previsto solo nel caso di effettivo cambio di abitazione*.

[Lit: The owners of motor vehicles and the owners of driving licence are not obliged to change their address on the car registration document and driving licence, *because the duty only exists in case of actual dwelling change*.]

c) Le ricordiamo inoltre che, ***quando** un immobile viene dichiarato inagibile o inabitabile*, bisogna presentare la denuncia di variazione I.C.I. prevista dall'art. 10, comma 4, del Decreto Legislativo 504/92.

[Lit: We remind you that, *when a real estate is being declared unlivable or uninhabitable*, you must report the I.C.I. variation, as it is established under article 10, paragraph 4, of Decree Law 504/92.]

(5)

(a) La medesima circolare ministeriale suggerisce altresì che il Comune, ***allo scopo di** evitare contestazioni* che potrebbero comportare il ritiro dei documenti [...]

[Lit: The same ministerial circular also suggests that the Municipality, *in order to avoid reprimands which may entail the revocation of the documents* […]]

(b) La variazione anagrafica in esame non comporta per i proprietari di autoveicoli e per i titolari di patente di guida l'obbligo di fare aggiornare la carta di circolazione e la patente di guida, ***in quanto** tale obbligo è previsto dal Codice della Strada soltanto per i casi di trasferimento effettivo di abitazione.*

[Lit: The examined registry variation does not require the owners of motor vehicles and the owners of driver license to update the vehicle registration and the driving license, in that this duty is established by the Highway Code only in cases of effective change of residence.]

(c) Si ricorda che, *mantenendo il regime del diritto di superficie*, qualunque passaggio di proprietà, affitto, cambio societario, ecc. dovrà essere autorizzato dal Comune di Schio [...]

[Lit: It is remembered that, ***holding*** *the leasehold regime*, any transfer of property, rental, corporate change etc. will have to be authorized by the Municipality of Schio]

If we assume this interpretation to be on the right track, the statistical values reporting a lower vs. higher distribution of subordinate conjunctions among the two corpora are not accidental but, instead, seem to reflect both a peculiarity of *bureaucratese* and a possible simplification strategy, which can be adopted to better explain the logical links and the sequence of events in highly informative texts, so that reducing the inference load and enhancing the reader's comprehension. Such considerations, if clearly suggest the need of refining the theoretical paradigm of subordination as equivalent to complexity, make it crucial to inspect more in-depth the output of the morpho-syntactic and syntactic annotation layers.

3.6.2.3 Verbal inflections and the distributions of pronouns

As observed in paragraph 3.2, Italian *bureaucratese* shows a peculiar morpho-syntactic tendency with respect to the large use of implicit clauses, *i.e.* propositions headed by nonfinite verbal mood (*i.e.* participles, gerunds and infinitives), less attested in ordinary usage language but instead typical of legal drafting. However, differently from what occurs at lexical level, where a certain degree of specialization is sometimes compulsory, these structures have been interpreted as one of those linguistic devices that allow the public administration to add obscurity to its texts, for instance by omitting (in case of gerundives and infinitives propositions), or unmarking, the agent role of a sentence. As a consequence, the simplification guidelines insist on an appropriate use of verbal morphology in bureaucratic prose[129].

[129] See, for instance, the indication from the *Guide for drafting administrative acts* «to avoid implicit verbal morphology, such as gerundives and participials, whenever the corresponding explicit forms could be used», p.23.

Interestingly, the comparative distribution of the indefinite verbal forms between *Bur_orig* and *Bur_simp* turned out to be significant in giving quantitative support to these qualitative claims. In particular, the figure below (Fig. 12), which illustrates the distribution of **participial verbs**, makes it possible to confirm the highest representativeness of this verbal mood in the bureaucratic prose, but also the different distribution between the corpora (Bur_simp: 26.09 Bur_orig: 29.79 t=-2.148451, p<.05).

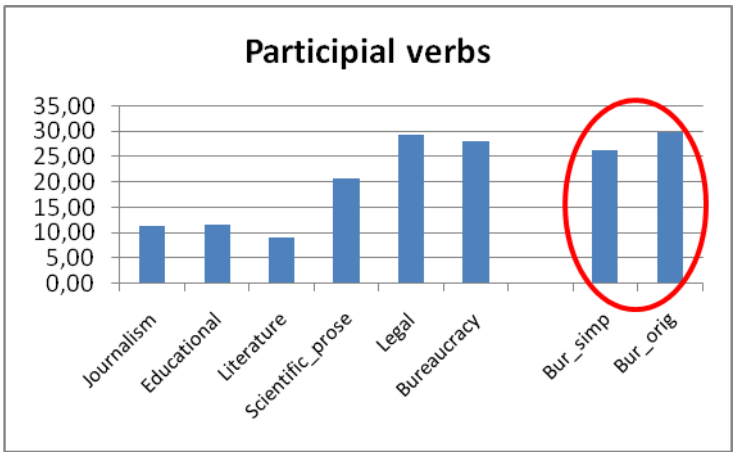


Figure 12: Participial verbs distribution across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	*% Participial Verbs*	
Journalism	Due Par	5.77	11.24
	La Rep	16.70	
Educational	Edu_child	11.39	11.63
	Edu_adult	11.87	
Literature	Lit_child	8.78	9.12
	Lit_adult	9.45	
Scientific_prose	Wiki	18.75	20.56
	Scient_art	22.37	
Legal language	It_Const	19.71	29.28
	Adm_acts	38.85	
Bureaucracy	Bur_simp	26.09	27.94
	Bur_orig	29.79	

Table 17: Percentage values of participial verbs across all the corpora.

However, these data are not completely straightforward, as in the Italian language past participles not only identify an indefinite mood, but they are also used in periphrastic forms with an auxiliary verb ("to be", "to have", "to get") to create the compound past tenses of finite moods (*e.g.* indicative present perfect). Thus, in order to better infer the 'actual'

function of participial verbs in our two corpora, we also calculated the percentage of participles that do not directly depend on auxiliary verbs.

Figure 13 reports the outcome of this analysis and seems to confirm that *bureaucratese* tends to abuse of elliptical constructs, headed by a participial verb, which are generally avoided in the corresponding rewritings and also very poorly attested in everyday language (see the average value of the journalistic genre, 10.16).

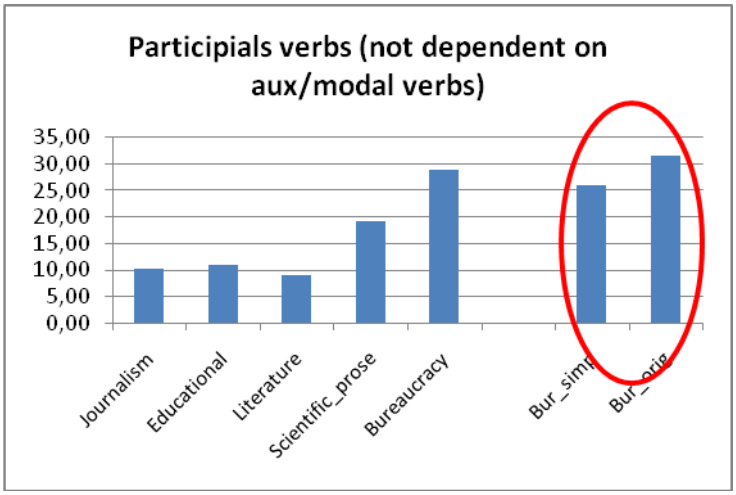


Figure 13: Participial verbs in elliptical constructions across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	"True" participial verbs (not dependent from aux/mod verbs)	
Journalism	Due Par	5.74	10.16
	La Rep	14.57	
Educational	Edu_child	10.25	10.87
	Edu_adult	11.50	
Literature	Lit_child	8.62	9.07
	Lit_adult	9.51	
Scientific_prose	Wiki	17.94	19.26
	Scient_art	20.59	
Bureaucracy	Bur_simp	26.02	28.81
	Bur_orig	31.69	

Table 18: Percentage values of "true" participial verbs across all the corpora.

Furthermore, within the comparative analysis of the verb inflectional paradigm, an additional hint towards the characterization of what makes a bureaucratic text more readable was provided by the analysis of the statistical distribution concerning the person feature in verbs[130].

[130] In Italian verb inflectional morphology, each tense for each finite mood has six voices: first, second and third singular and first, second and third plural.

Genre	Corpus	1 Sing	2 Sing	3 Sing	1 Pl	2 Pl	3 Pl
Journalism	Due Par	3.60	0.27	20.31	3.10	0.01	21.31
	La Rep	1.62	0.64	29.60	1.01	0.12	9.51
Educational	Edu_child	1.83	0.64	35.14	0.63	0.29	18.36
	Edu_adult	1.22	0.73	38.04	2.14	0.07	9.58
Literature	Lit_child	3.77	2.40	38.54	2.16	1.09	12.52
	Lit_adult	3.62	1.55	39.09	1.77	0.36	7.84
Scientific_prose	Wiki	0.42	0.88	32.34	0.18	0.02	16.64
	Scient_art	0.43	0.47	26.73	1.49	0.03	12.09
Legal language	It_Const	0.19	0.39	49.71	0.00	0.00	23.30
	Adm_acts	2.08	0.63	40.84	0.01	0.14	26.89
Bureaucracy	Bur_simp	2.32	0.44	16.78	**13.57**	0.50	3.83
	Bur_orig	2.77	0.84	24.64	**1.48**	0.01	5.20

Table 19: Percentage distribution of person feature in verbs across all the corpora.

Trying to paraphrase in a qualitative way the numerical scores reported in Table 19 (illustrated in Figure 14) we might say that "a clear public administration is made of people, thus speaks in first person plural". There is indeed outstanding statistical evidence making evidence the different distribution of this verb person in the two corpora (*But_orig*: 1.48; *Bur_Simp*: 13.57; Average: 7.52, T-value is 7.728236. $p < 0.01$.).

Conversely, what characterizes the original bureaucratic texts is a sharp preference for third singular verbs. On one side, the latter form agrees with the typical subjects of administrative documents, i.e. singular abstract nouns such as "this administration", "this office" and so on.; on the other, it also represents the verb form that we find in **impersonal sentences**, which are one of the major marker of *bureaucratese* syntax (cf. 3.2).

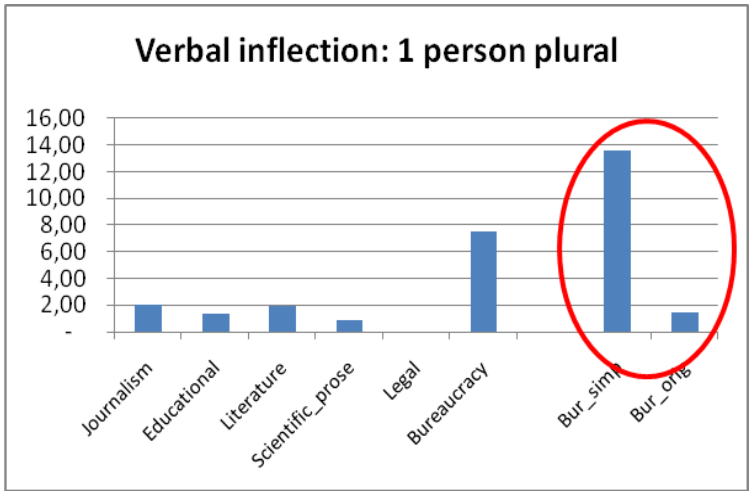


Figure 14: First person plural across major genres, *i.e.* simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Such a tendency towards impersonality can also be further elaborated by inspecting the internal distribution of pronouns, and more specifically, the occurrences of **personal vs. clitic pronouns.** As table 20 shows, if the former predominate in *Bur_simp*, the latter are the most instantiated pronominal class in *Bur_orig*.

Genre	Corpus	Personal Pronouns		Clitic Pronouns	
Journalism	Due Par	0.08		0.96	
	La Rep	0.21	0.14	1.59	1.28
Educational	Edu_child	0.72		2.23	
	Edu_adult	0.51	0.61	2.10	2.17
Literature	Lit_child	0.84		3.62	
	Lit_adult	0.66	0.75	3.15	3.38
Scientific_prose	Wiki	0.15		1.12	
	Scient_art	0.12	0.14	1.05	2.14
Legal language	It_Const	0.13		0.96	
	Adm_acts	0.08	0.11	0.42	0.69
Bureaucracy	Bur_simp	0.55		1.21	
	Bur_orig	0.19	0.37	1.76	1.48

Table 20: Percentage distribution of personal and clitic pronouns across all the corpora

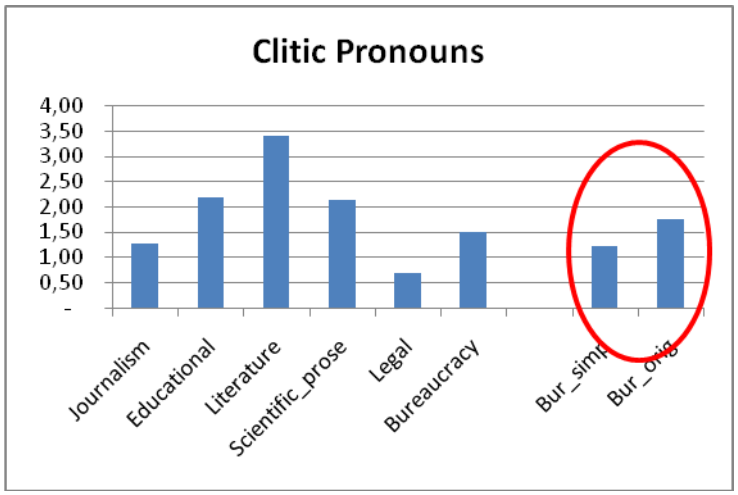


Figure 15: Clitic pronouns across major genres, *i.e.* simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

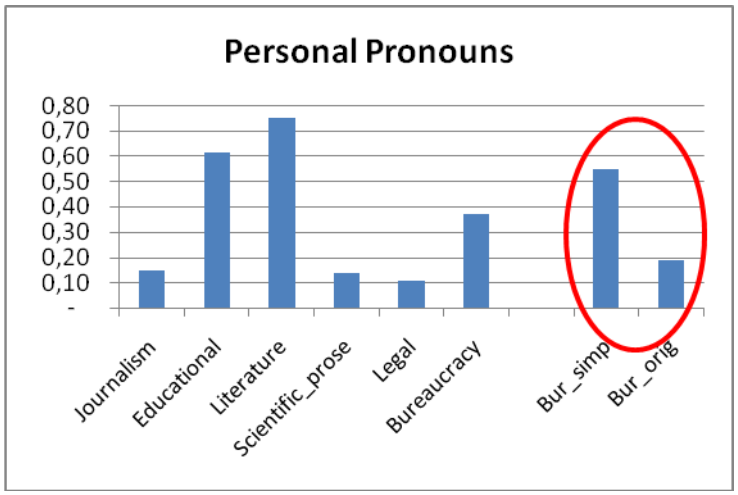


Figure 16: Personal pronouns across major genres, *i.e.* simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

However, this is a finding that requires a more in depth consideration. While a large use of personal pronouns is a symptom of a more reader-focused style, the "raw" value measuring the distribution of clitics is not sufficient itself to signal the presence of impersonal structures. This is because Italian clitics can play different syntactic roles (accusative, dative, partitive, reflexive), with the only clitic '*Si'* used to derive impersonal constructions. Thus, in order to disambiguate the grammatical function of the clitic pronoun, we have cross-checked the morpho-syntactic output with the output of the syntactic annotation, which allows distinguishing the relationship between: i) a clitic pronoun and a verbal head used in pronominal form (i.e. marked by a "clit" arc) and ii) the relationship between a verbal head and an accusative or dative clitic (marked, respectively, as an object ("obj") and indirect complement ("comp_ind") dependency). In line with the expectations, the result has confirmed a significant difference with respect to the distribution of **clitic dependencies**, which are higher in the original bureaucratic corpus and very low in the simplified counterpart (Fig. 17) [131].

[131] However, it is worth pointing out that for some varieties of regional Italian, e.g. Tuscan dialect, the clitic "si" has the force of a first personal plural, thus it might not be perceived by the reader as an impersonal form.

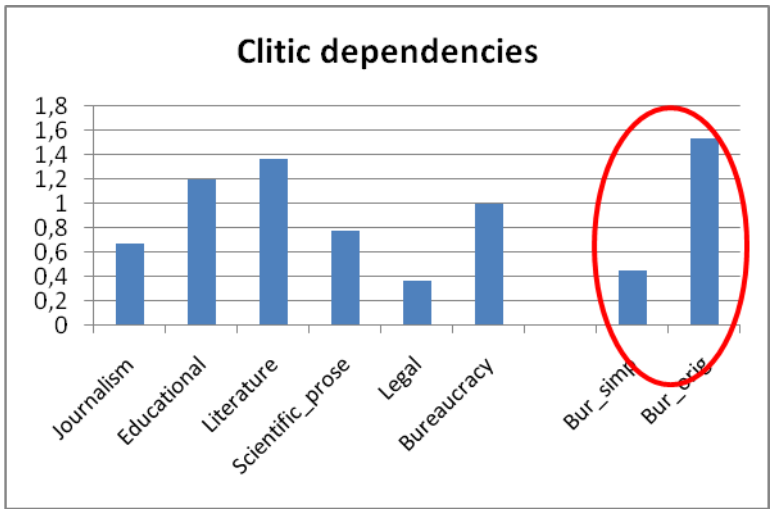


Figure 17: The distribution of clitic dependencies across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus

Genre	Corpus	Clitic dependencies	
Journalism	Due Par	0,48	0.67
	La Rep	0,86	
Educational	Edu_child	1,19	1.19
	Edu_adult	1,19	
Literature	Lit_child	1,52	1.36
	Lit_adult	1,20	
Scientific_prose	Wiki	0,81	0.77
	Scient_art	0,73	
Legal language	It_Const	0,47	0.37
	Adm_acts	0,27	
Bureaucracy	Bur_simp	0,45	0.99
	Bur_orig	1,53	

Table 21: Clitic dependencies across all the corpora

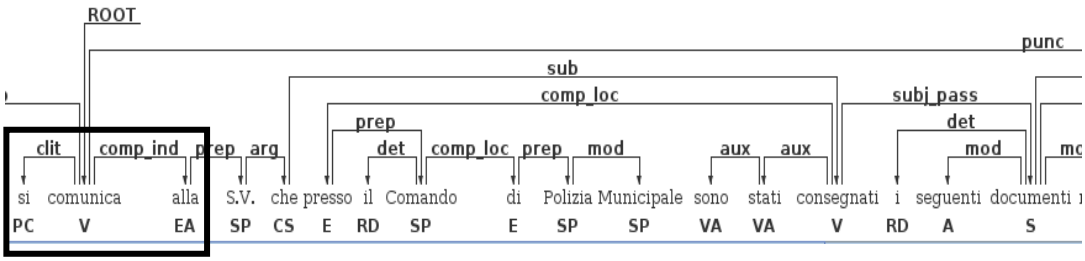
Before concluding this section, it is worth remarking that all the features so far discussed (person morphology, clitic pronouns and clitic dependencies) are not to be viewed as implicated in a "general-oriented" notion of readability; they should instead represent the corresponding operationalization, from a computational linguistics perspective, of different stylistic choices that, taken as a whole, allow us to capture "signatures" of the process of rewriting inspired by the bureaucratic language simplification guidelines[132].

[132] Another clarifying example of this correlation is taken again from the *30 Rules for writing clear administrative acts* (Cf. 3.2.1) [translation mine]: «What can be used to replace the impersonal? A personal form that agrees with a subject expressing the name of the office or that of the administration, but also a first person plural verbal form, without any subject's reference, such as "vi trasmettiamo" [we transmit you], "vi informiamo" [we inform you], "abbiamo respinto la richiesta" [we have rejected the request]). The latter solution [allowed in that Italian is a null subject language] combines the need of not highlighting the writer, in that he/she is writing not in his/her name but in the name of the administration, to the choice of using a direct and common form, like the first plural».

To give an example, the following sentences taken respectively from *Bur_orig* (6) and *Bur_simp* (7), illustrate the way an impersonal sentence has been usually treated in the rewriting process. By looking at the different syntactic annotation (in particular the label marked by the square), it is also showed how these changes can be intercepted within the output of the parsed sentence.

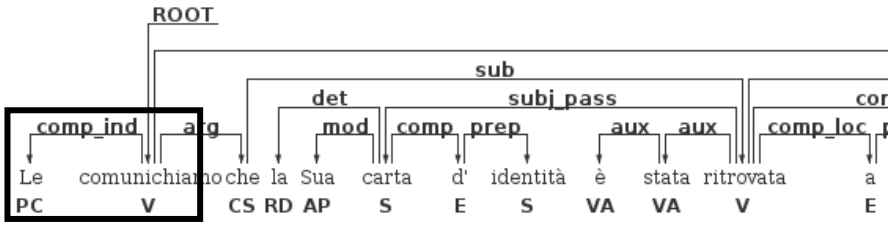
(6)

Si comunica alla S.V. che presso il Comando di Polizia Municipale sono stati consegnati i seguenti documenti [...]
[*S.V.*[133] *is informed* that the following documents have been delivered to the Local Police […]



(7)

Le comunichiamo che la Sua carta d'identità è stata ritrovata [...]
[*We inform you* that your Identity Card has been found […]



[133] S.V. is the acronym of 'Signoria Vostra', almost impossible to translate in English. It is an archaic and very formal pronoun of address, whose usage is restricted to bureaucratic communication.

3.6.2.4 Parse tree features

Up to now, we have tried to account for a characterization of the impersonal style traditionally used in official writing by examining the distribution of fine-grained morpho-syntactic features (specifically, verb inflections and pronouns) and comparing them to the percentage values attested in the simplified rewritings.

The last part of this section will be devoted to showing how the simplification process has also contributed to modify those syntactic properties that have already proven to be involved in assessing the readability level of general-purpose texts. The properties involved are captured by the following parameters: the **parse tree depth**; the **length of the dependency links** and the distribution of **complex** (*i.e.* recursively embedded) **prepositional complements by depth**.

As we could expect, we observe that the manual process of rewriting has affected all of these features, yielding to: *i)* lower parse trees (Fig. 18); *ii)* shorter dependency links (Fig. 19 and Fig 20) and *iii)* a lower frequency of deeper embedded prepositional chains (Tab. 25). These tendencies in turn confirm the effectiveness of such selected features to identify the typical "clumsy" syntax affecting *bureaucratese*.

As an example, let's consider the feature measuring the length of the dependency links, which is here calculated in terms of the average number of tokens between a head and its dependent. Within the sub-corpus of the original bureaucratic texts under examination, such a parameter allows us to infer the abuse of asides and parenthetical clauses (see, *e.g.*, sentence (8)), which is a typical pattern of *bureaucratese* syntactic structure (cf. §3.2). As we know from psycholinguistic evidence (§1.3), constructs of this kind overload sentence processing: not only they increase the overall sentence length but they also interrupt the flow of information, especially when occur between the subject and the main verb or between the verb and its complements. Thus in the attempt of simplifying these constructs, the author of the rewriting might have decided either to split the original sentence containing a parenthetical clause into two independent sentences or to reformulate the original sentence so that to preserve the adjacency requirements between the head and its dependent, maintaining the whole content or dropping irrelevant parts; this is what occurred, e.g., in sentence (9), which is the rewriting of sentence (8) from *Bur_Orig*.

(8) Si comunica che, *a seguito della Vostra richiesta di poter realizzare la manifestazione indicata in oggetto*, l'Amministrazione Comunale con argomento di Giunta nr. 99 del

23.03.04, ha espresso parere favorevole allo svolgimento della stessa in Piazza Europa per Domenica 9 maggio c.a.

[Lit: It is informed you that, *following your request to hold the aforementioned manifestation*, the municipal administration via the Municipial Board decision nr. 99/ 23.03.04, has delivered a favorable opinion of holding the aforementioned one in Piazza Europa on Sunday, May 9.]

(9) Vi comunichiamo che è stata accolta la vostra richiesta di svolgere la IX edizione di "Bimbi in piazza" per domenica 9 maggio 2004 in Piazza Europa. Vi invitiamo pertanto a contattarci per gli adempimenti amministrativi, tecnici e logistici.

[Lit: We inform you that it has been accepted your request of holding the IX edition of "Bimbi in piazza" on Sunday, 9 may 2004 in Piazza Europa. We thus invite you to contact us for the administrative, technical and logistical fulfillments.]

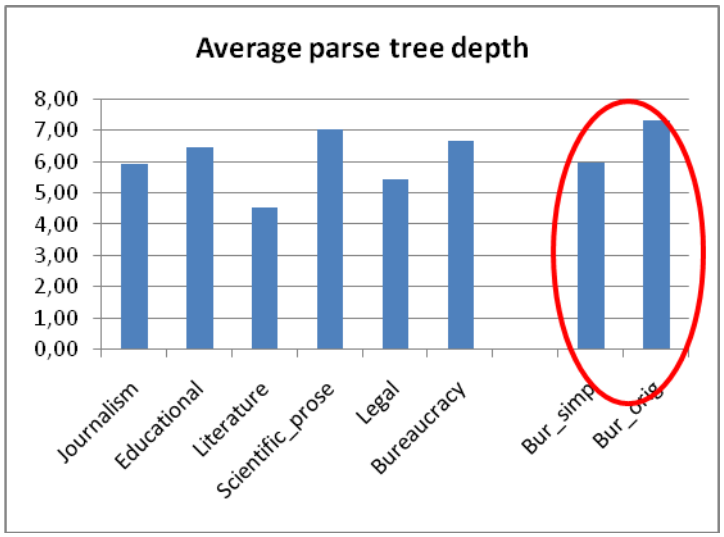


Figure 18: Average parse tree depth across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus

Genre	Corpus	Average parse tree depth	
Journalism	Due Par	5.29	5.90
	La Rep	6.51	
Educational	Edu_child	5.54	6.45
	Edu_adult	7.36	
Literature	Lit_child	4.51	4.54
	Lit_adult	4.57	
Scientific_prose	Wiki	6.47	7.04
	Scient_art	7.62	
Legal language	It_Const	4.73	5.44
	Adm_acts	6.15	
Bureaucracy	Bur_simp	5.96	6.64
	Bur_orig	7.32	

Table 22: Average parse tree depth across all the corpora

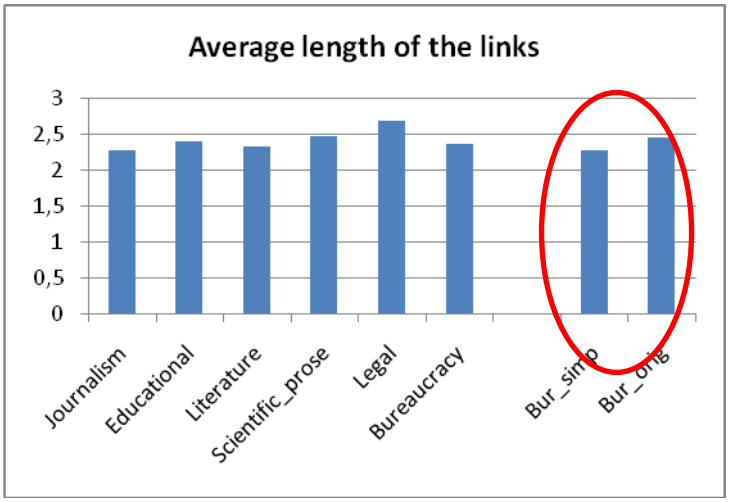


Figure 19: Average length of the links across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	Average length of the link	
Journalism	Due Par	2.16	2.27
	La Rep	2.39	
Educational	Edu_child	2.24	2.39
	Edu_adult	2.54	
Literature	Lit_child	2.25	2.33
	Lit_adult	2.40	
Scientific_prose	Wiki	2.46	2.47
	Scient_art	2.48	
Legal language	It_Const	2.34	2.68
	Adm_acts	3.03	
Bureaucracy	Bur_simp	2.26	2.36
	Bur_orig	2.45	

Table 23: Average length of the links across all the corpora.

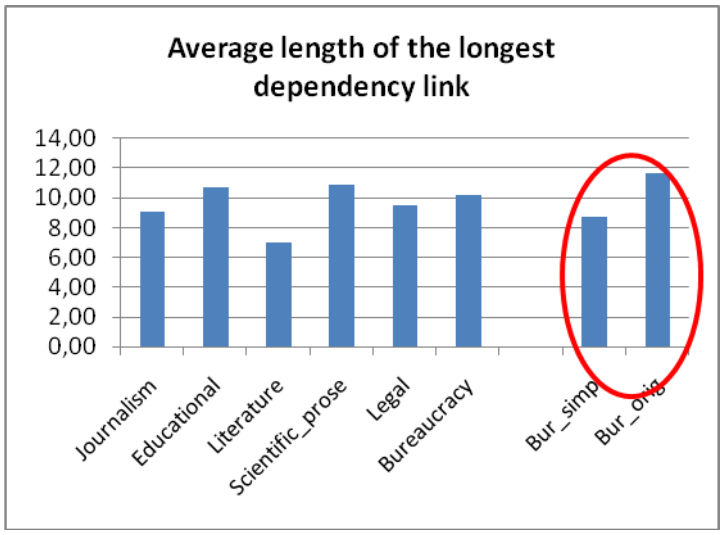


Figure 20: Average length of the longest link across major genres, i.e. simple/complex varieties collapsed, and with respect to the internal distinction of the bureaucratic corpus.

Genre	Corpus	Average length of the longest dependency link	
Journalism	Due Par	7.91	9.10
	La Rep	10.28	
Educational	Edu_child	8.89	10.69
	Edu_adult	12.50	
Literature	Lit_child	6.63	7.03
	Lit_adult	7.43	
Scientific_prose	Wiki	9.88	10.92
	Scient_art	11.96	
Legal language	It_Const	6.75	9.51
	Adm_acts	12.28	
Bureaucracy	Bur_simp	8.69	10.17
	Bur_orig	11.65	

Table 24: Average length of the longest dependency link across all the corpora.

Genre	Corpus	Distribution of prepositional chains by depth					
		1 embedded complement		2 embedded complements		≥ 3 embedded complements	
Journalism	Due Par	79.40	75.86	17.02	19.23	3.27	4.61
	La Rep	72.32		21.43		5.94	
Educational	Edu_child	81.46	78.59	15.73	17.65	1.17	2.24
	Edu_adult	75.72		19.57		3.31	
Literature	Lit_child	83.18	80.17	14.11	14.86	1.73	2.19
	Lit_adult	77.16		15.61		2.64	
Scientific_prose	Wiki	70,87	67.99	22.03	23.30	6.41	7.77
	Scient_art	65.11		24.57		9.12	
Legal language	It_Const	70.15	61.51	23.54	24.81	6.30	13.36
	Adm_acts	52.87		26.08		20.41	
Bureaucracy	Bur_simp	61.76	61.30	24.69	25.35	**10.41**	11.18
	Bur_orig	60.84		26.02		**11.94**	

Table 25: Distribution of prepositional chains by depth across all the corpora.

3.7 Relative clauses: a qualitative analysis

This section presents a manually carried investigation on the use of restrictive relative clauses within the bureaucratic corpus. Such a focused analysis into a well-known marker of sentence complexity (§1.3) has been prompted by some observations about the limits of the automatic analysis here adopted as the only mean to capture some high-leveled linguistic complexity cues..

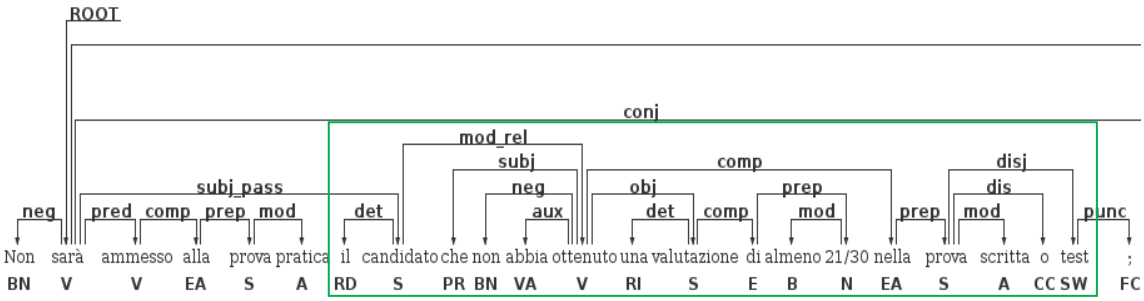
More specifically, among the features that it was possible to automatically extract from the syntactic dependency output, we inspected the use of subordinate clauses, both with respect to their proportion over the main clause and according to their internal structure (§3.6.1.6). However, we also noted that these data were underestimated (cf. footnote 127), since the corresponding parameters do not comprise the frequency of relative clauses (i.e. the dependency arc labeled as "mod_rel"). This is because the identification of the correct attachment side for relative clauses not only can pose difficulty to humans but also constitutes a typical domain of failure for statistical parsers (Siddharthan, 2002); such a consideration also holds for the dependency parser adopted for this study (i.e. DeSR parser, §2.6.1), which indeed obtains quite low accuracy scores with respect to the appropriate treatment of these clauses[134].

Moreover, while the distinction between (headed) Subject (SRC) and Object Relative clauses (ORC), despite not explicitly marked by different labels (i.e., there exists a unique "mod_rel" tag for identifying these sentences), can be yet derived by looking at the labeled arc (i.e. subj vs. obj) linking the relative pronoun to the embedded verb of the relative clause (see examples (10) and (11)), no other cue is provided about the realization of the latter as passive object relatives (PORs). Nevertheless all such distinctions are crucial in the attempt of providing cognitively-motivated metrics of syntactic complexity, as they have a different impact on locality issues and, ultimately, on sentence processing load (§1.3.4); thus, we believe that monitoring these sentences can offer a more sophisticated analysis of the degree of sentence complexity within texts.

(10) SRC: Non sarà ammesso alla prova pratica *il candidato che non abbia ottenuto una valutazione di almeno 21/30 nella prova scritta o test.*

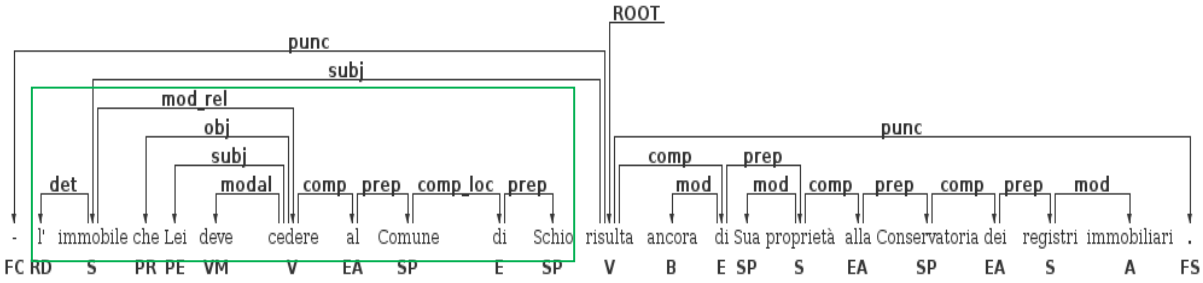
[134] In terms of Labeled Attachment Score (LAS): precision 53.03; recall 61.04 (f-measure: 56.91).

[Lit: It will not be admitted to the practice test *the candidate who has not received a rating of at least 21/30 in the written test.*]



(11) ORC: *L'immobile che Lei deve cedere al Comune di Schio* risulta ancora di Sua proprietà alla Conservatoria dei registri immobiliari.

[Lit: *The property that You have to transfer to the Municipality of Schio* is still in your ownership [...]]



An additional research question has motivated the investigation here proposed. As we did not find any explicit mention addressing the treatment of relative clauses as a type of post-nominal modification in the simplification guidelines for bureaucratic language, it is interesting to examine whether these constructs can contribute in distinguishing original vs. simplified administrative texts, not only for what concerns their overall distribution within the two corpora, but especially according to their classification into different typologies. With respect to this point, we designed a qualitative analysis inspired to the study of Belletti and Chesi (2011), who focused on different types of relative clauses within three corpora representative of standard spoken Italian (i.e. CHILDES, Siena University Treebank (SUT) and Corpus di Italiano Televisivo (CIT)).

3.7.1 Method

The corpus-based analysis of relative clauses aimed at distinguishing them into the following macro-typologies:

a) Subject-relative clauses (SRCs);
b) Object-relative clauses (ORCs);
b) Passive object relative clauses (PORs);
d) Indirect relative clauses (IORs).

The extraction of relative clauses from the two varieties of bureaucratic texts, i.e. the simple and the complex one, was carried out semi-automatically, so that to overcome the poor precision and sophistication of the automatic annotation in distinguishing the different typologies of relative clauses we were interested in. More specifically, following Belletti and Chesi (2011)'s considerations, we relied on a regular expression like the one suggested in their work (here repeated in 12), which allowed us to retrieve all sentences containing the token "che". Clearly, as the authors noted, the outcome is not precise, since "che" in Italian is used both as a relative pronoun in (non-reduced) SRCs/ORCs/PORs and as a complementizer in declarative clauses; however, this approach offers a method to discriminate the actual relative clauses out of the whole set of sentences, thus reducing the quantity of data to be manually examined. The results of this preliminary analysis are provided in Table 26 (and graphically in Figure 21).

(12) Regular expressions using "grep":
grep -i -n -E
"TIER:([[:space:]]|[[:punct:]]|[[:alpha:]])*[[:space:]]che[[:space:]]"

Relative clauses typologies	**Bur_Orig**		**Bur_Simp**	
	N°	**%**	**N°**	**%**
RS	67	51.4	37	45.68
RO	6	4.62	11	13.58
POR*	42	32.31	27	33.33
IOR	15	11.54	6	7.41
Tot	130	100	81	100

Table 26: RC macro-classes with SRs split in active (SRs) and passive (PORs) SRs.

(* For what concerns PORs, these data include both full and reduced (long) relatives.)

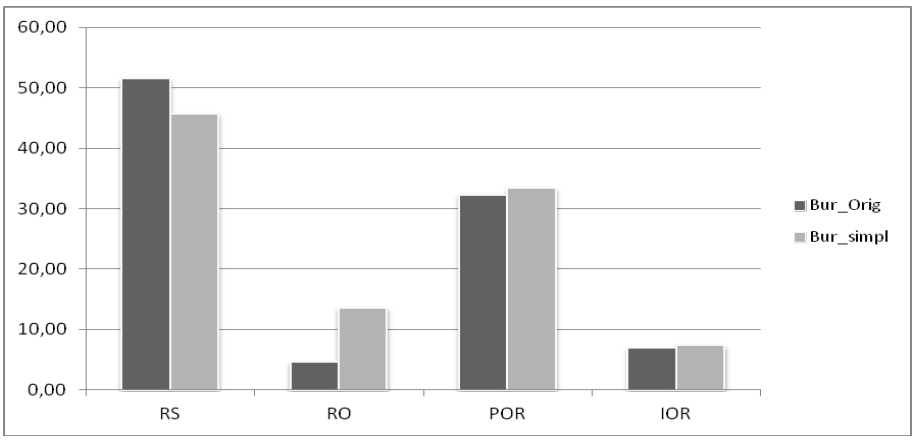


Figure 21: RC macro-classes with SRs split in active (SRs) and passive (PORs) SRs.

An overall look at the table reveals an almost twice distribution of relative clauses, independently from their internal classification, in the original version of the bureaucratic corpus with respect to the simplified counterpart: such a finding signals that the manual process of re-writing has taken into account these particular complex sentences, trying to restate the meaning conveyed by the relative clause in a different manner.

Interestingly, whereas SRCs are considerably higher than ORCs in both the corpora – an evidence that is in line with data from elicited production and natural spoken language – we can observe an unexpected, despite very slightly marked, higher frequency of the latter in the simplified texts with respect to *Bur_Orig*. However, none of the ORCs occurring in the simplified texts instantiate the syntactic configuration that is more problematic according to the featural approach to syntactic locality (§1.3.4); that is to say, in none of them, the head of the relative clause and the intervening subject share the lexical restriction feature. Instead they all realize the target (i.e. the relative head) as an inanimate full lexical NPs and the intervener (i.e. the embedded subject) either as an explicit (cf. (13), (14)) or a null pronoun (cf. (15),(16)).

(13) ***I certificati*** *che* ***Lei*** *ha richiesto* sono soggetti al tributo di bollo.

[Lit: The certificates that *You* have requested are subject to stamp duty.]

(14) ***L'immobile*** *che* ***Lei*** *deve cedere* al Comune di Schio risulta ancora di sua proprietà.

[Lit: The property that You have to transfer to the Municipality of Schio is still in your ownership]

(15) ***La convenzione preliminare*** *che* $_{pro}$ *avete sottoscritto* vi obbliga [...]

[Lit: The preliminary agreement that (you) have signed forces you […]]

(16) Dovrai risarcire ***i danni*** *che eventualmente* $_{pro}$ *procurerai.*

[Lit: (you) will have to refund the damages that (you) may cause]

The second finding that has deserved a more in depth analysis concerns the wide representativeness of passive object relatives (PORs) in both corpora. Although they were present within the corpora of standard Italian investigated by Belletti and Chesi[135], such a typology of relative clause in the bureaucratic corpus considered in our study was consistently more attested.

Clearly, the different modalities, i.e. spoken vs. written, as well as the diverse language variety taken into account, i.e. standard language vs. bureaucratic language, do not allow us to make a comparative analysis; nevertheless, it is interesting to observe that PORs, from the point of view of processing, are easier than active object ORCs, since the different underlying computation (§1.3.4, spec. footnote 46) enables the parser to overcome the problem of the intervener; thus, our quantitative results could be taken as a feature signalling "simplicity" within a typology of complex texts.

However, we need to take into account that (both full and reduced) PORs can be also realized in a short form, namely omitting the "by-phrase", which is the thematic subject of the verb within the relative clause. We might speculate that it is exactly this property that can potentially make these structures less adequate from the perspective of written discourse, and especially within highly constrained texts (Sabatini, 1990) like the bureaucratic documents, which should reduce to the most extent the "set" of possible interpretations set forth by the receivers; this purpose can be achieved, e.g., if the writers avoid omitting necessary information (when it is not clearly retrievable from the context), such as the by-phrase making explicit the identity of the agent (e.g. the authority requiring a specific action, prescribing a rule and so on).

With the aim of better characterizing PORs within the corpora, we thus refined the first analysis by also including these cases, namely introducing the following two sub-categories of PORs:

[135] PORs (both full and reduced) are attested at 13% in CIT; 10% in CHILDES (adult); 13% in SUT.

- POR long: they include both full (i.e. explicit (thematic) subject and explicit relative pronoun), (example (17)) and reduced (i.e. explicit (thematic) subject and implicit relative pronoun) PORs, (example (18));

- POR short: they include both full (i.e. implicit (thematic) subject and explicit relative pronoun) (example (19)), and reduced (i.e. implicit (thematic) subject and implicit relative pronoun) PORs, (example (20));

(17) […] secondo *le disposizioni che saranno impartite dal coordinatore*.
[Lit: according to the instructions that will be provided by the director]
(18) Abbiamo trasmesso *la richiesta presentata dal dott. Mario Rossi* […]
[Lit: We transmitted the request advanced by Dr. Mario Rossi]
(19) Sarà formulata *una graduatoria che potrà essere resa pubblica*.
[Lit: It will be formulated the ranking that should be published]
(20) *I dati conferiti* possono essere comunicati […]
[Lit: The provided data can be transmitted]

Table 27 (and Figure 22) illustrates the findings of this second analysis.

Relative clauses typologies	**Bur_Orig**		**Bur_Simp**	
	N°	**%**	**N°**	**%**
RS	67	30,87	37	29,37
RO	6	2,76	11	8,73
POR_long	42	19,35	27	21,43
POR_short	87	40,09	45	35,71
IOR	15	6,91	6	4,76

Table 27: RC macro-classes with SRs split in active (SRs) and passive (PORs) SRs. PORs are distinguished in turn into long and short form.

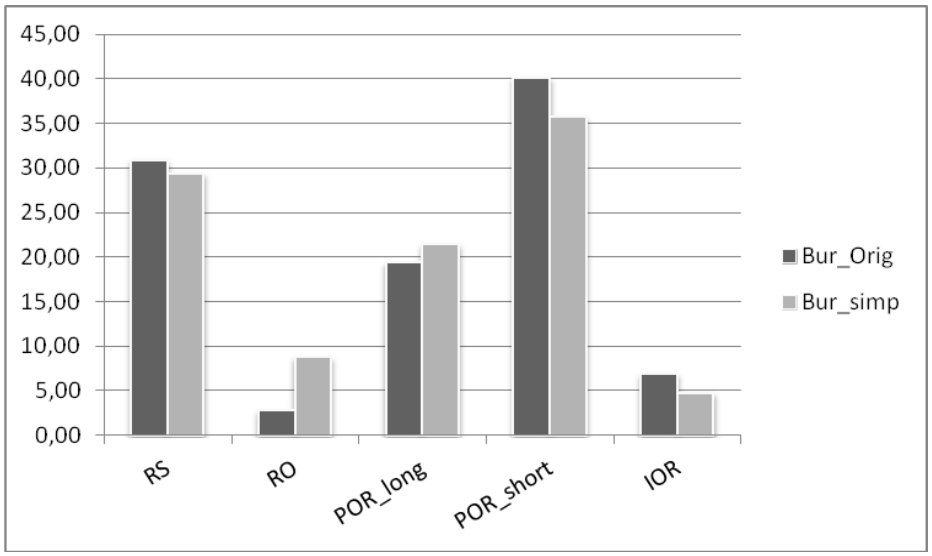


Figure 22: RC macro-classes with SRs split in active (SRs) and passive (POR_long; POR_short) SRs.

Although the occurrence of PORs in the short form keeps being significant within the corpus of simplified texts, we registered a reduction of 5 percentage points (40.09 vs. 35.71) with respect to the corpus of original bureaucratic texts. Without neglecting the many of the reduced PORs in the short form attested in the corpora look like simple form of adjectival modification, this finding might corroborate the hypothesis that the author of the rewriting paid particular attention to make explicit the agent role as a way to make the text more comprehensible.

3.8 Summary

In this chapter we have conducted a comparative linguistic profiling investigation of a corpus of Italian bureaucratic texts from a computational linguistic perspective, focused on the assessment of a set of multi-leveled linguistic features automatically extracted from text. On the basis of the patterns detected – which in most cases complied with what already established by traditional descriptive analyses – we believe that these features do indeed contribute to distinguish "genre-complexity" markers from stylistic signals of *bureaucratese*, thus proving the efficacy of the adopted perspective.

However, some remarks have to be pointed out. First, it seems quite necessary to validate the observations derived by this study on a larger "parallel corpus", and even better, on a corpus balanced in terms of the internal representativeness of diverse typologies of administrative texts, since criteria such as the intended receiver (external vs. internal) and the

juridical value of the document (§ 3.3)) might affect the extent to which a bureaucratic text can be rewritten to accomplish standards of language clarity and simplicity.

Second, some of the features deemed as factors of linguistic complexity in general-purpose texts – such as the low percentage of words belonging to a "basic" vocabulary of a language – are not particularly predictive when the purpose is to assess the readability of texts showing a certain degree of technicality and specialization like the bureaucratic ones. It is not surprising indeed that these texts turn out to contain unknown or less familiar words: what is crucial is to distinguish the genuine "technicisms", i.e. words very difficult to discard or replace, from the "pseudo-technicisms", namely unnecessary rare words only use for stylistic purposes. The identification of a vocabulary of domain terminology is another research area that can be tackled by relying on automatic term recognition methods, which have already been tested with promising results for Italian corpora of legal language, and also with respect to the extraction of multi-word expression (Bonin *et al.*, 2010).

Third, while many formal metrics informed by linguistic and psycholinguistic research can be operationalized with a substantial margin of reliability, especially for the syntactic domain (e.g. parse tree depth, subordination, number of dependents for verbal head), there is still room for improvement. As a great body of empirical data shows, syntactic structures display subtle properties that are responsible of making the sentence harder to comprehend: it is the case of the intervention effects in moved-derived dependencies (e.g. object relative clauses), which clearly cannot be captured by relying on a "distance-based" parameter calculated in terms of words; their assessment, particularly when a text is intended for low-skilled or language-impaired readers, makes it necessary to support automatic linguistic analysis with more focused, qualitatively carried, inspections. This issue is of particular importance when we attempt to automatize a method to gauge linguistic complexity at sentence level, and not only at text level, which, as we will see in the next Chapter, is the prerequisite of automatic text simplification research.

Chapter 4

From readability assessment to text simplification: an exploratory study for the Italian language

4.1 Introduction

The need of recognizing and operationalizing linguistic complexity markers in natural language data is a methodological precondition for large-scale investigations not only limited to the analysis of texts from readability assessment purposes. NLP-based metrics of linguistic complexity have proven valuable research tools in a variety of fields of applied linguistics, such as:

- child language acquisition, e.g. with the aim of measuring syntactic complexity of children's oral productions (Sagae *et al.*, 2005) or comparing the development of oral and written language skills (e.g. Silva *et al.*, 2010);
- adult language impairment, for measuring the impact of neurodegenerative impairment on speech and language (Roark et al., 2007; Sahakian and Snyder, 2012; Fraser, 2014);
- text simplification, for evaluating the difficulty of text as a preliminary step for its automatic simplification (Siddhartan, 2014; Saggion, 2014 for a survey).

This chapter is concerned with the last scenario and it presents a first study that has tackled the issue of building the necessary resources and methods for investigating text simplification in Italian, a language for which this task is still largely under-searched[136].

It has to be noticed that Automatic Text Simplification (ATS) is related to computational readability assessment research in many ways. From a functional viewpoint, the availability of "advanced" readability measures capable of assessing the difficulty of texts, not only with respect to their global structure but especially at sentence level, can be considered as a first step towards their automatic simplification. Indeed, not only a sentence-level readability score gives a more accurate metric for identifying which sentences or subparts of them need

[136] The following discussion is based on work published in Brunato *et al*. (2015).

to be simplified (Dell'Orletta *et al*., 2011; Aranzabe *et al*., 2013) but it can also be used to assess whether the simplification has actually provided an easier text (Štajner and Saggion, 2013).

On the methodological side, the two tasks also show important analogies. Similarly to the current trend in readability assessment, recent approaches to automatic text simplification tend to rely more and more on data-driven algorithms. While in the context of readability evaluation, the "gold" corpora used for training are typically collected so that to represent different readability levels, for what concerns ATS research, the availability of parallel monolingual corpora provides a considerable advantage. Such a kind of resources comprises the original and the simplified version of the same text, typically aligned at sentence-level either manually or automatically, and they offer the opportunity to investigate the real editing operations that human "simplifiers" (e.g. linguists, teachers) perform on a text, as well as their computational treatability.

This chapter is organized as follows; paragraph 4.2 outlines a state of the art in automatic text simplification; in section 4.3 we will report the preliminary outcomes of ongoing work for the Italian language.

4.2 State of the Art

Automatic Text Simplification (ATS) has been defined as the process aiming at reducing lexical and syntactic complexity of a text by preserving its original meaning and information content (Siddhartan, 2014). Although this task shares many similarities with text summarization and sentence compression, the given definition highlights the different nature of ATS, in which all information should be maintained, while text summarization usually drops unnecessary data from the input text.

ATS is a relatively novel field of research in NLP community but is receiving growing attention over the last few years, due to the implications it has for both machine- and human-oriented tasks. For what concerns the former, ATS has been exploited as a preprocessing step, which provides an input that is easier to be analyzed by other NLP modules, so that to improve the efficiency of those tasks relying on linguistic analysis, e.g., parsing, machine translation and information extraction. With respect to the latter, ATS can play a crucial role in supporting the development of educational and assistive technologies.

Under this perspective, computational methods for simplifying texts have been proposed to create contents more adapted to the needs of particular readership, like children (De Belder and Moens, 2010), L2 learners (Petersen and Ostendorf, 2007), people with low literacy skills (Aluísio *et al.*, 2008), cognitive disabilities (Bott and Saggion, 2014) or language impairments, such as aphasia (Carroll *et al.*, 1998) or deafness (Inui *et al.*, 2003).

Despite the different applicative goal, a full ATS architecture needs to cope with both the aspects of text analysis and text (re)generation. Namely, there is a first stage, which is devoted to analyzing an input text and detecting the areas of complexity on it by focusing at the level of sentence, and a subsequent stage, in which the complex sentence is rewritten into a simpler version. In a semi-automatic ATS system, only the first aspect is fully automated but the system is able to provide some possible re-editing suggestions, letting the final user to decide how to modify the sentence.

To date, the creation of both full and semi-automatic ATS systems has been addressed by three main approaches. The more traditional one is based on the use of hand-crafted rules conceived to simplify a predefined set of constructs, typically at syntactic level, which are deemed proxies of complexity both for humans and automatic parsers. In this context, special attention has been paid to relative clauses, subordinate clauses, appositions, passive sentences, whose structure is typically identified from the output of a syntactic parser and simplified according to *ad hoc* rules. This is the approach followed by the *Practical Simplification of English Texts* (PSET) project (Carroll *et al.*, 1998), which was the first large research undertaking ATS for the benefits of human readers, in this case English-mother tongue adult aphasics. Beyond the syntactic simplification rules, this project also developed a lexical simplification module, based on the use of WordNet (§2.3.1.1), to replace complex terms with easier synonyms.

As anticipated above, in more recent years the availability of larger parallel corpora, such as the English (EW) and Simple English Wikipedia (SEW)[137], has opened up the possibility for a more consistent use of machine learning algorithms for automatically acquiring simplification rules. This is the approach followed by e.g. Woodsend and Lapata (2011), who based their ATS system on a quasi-synchronous grammar, Zhu et al. (2010), who adapted a Statistical Machine Translation (SMT) algorithm to implement simplification operations on the parse tree, and Narayan and Gardent (2014), who similarly adopted SMT techniques but also combined a deep semantic representation of the sentence.

[137] http://simple.wikipedia.org/wiki/Simple_English_Wikipedia [last access: 01/07/2015]

It is worth noting that both hand-written and automatically acquired rules have advantages and shortcomings. While the former can potentially account for the maximum linguistic information, they are extremely costly to develop and consequently they tend to cover only few lexical and syntactic patterns; on the other side, data-driven approaches require the least linguistic knowledge but they are not feasible to carry out without a large quantity of aligned data. A possible alternative seems to be offered by "hybrid" approaches, such as that described by Siddharthan and Angrosh (2014), in which a combination of automatically acquired lexical rules and hand-crafted syntactic rules outperformed the state of the art.

However, all these systems exploit the Wikipedia dataset as a training corpus. Resources of this kind are lacking, or much less restricted, for languages other than English, making it rather impossible to approach ATS as pure machine learning task. To cope with this issue, in many cases parallel monolingual corpora have been manually built or hand-aligned from existing resources: these corpora, which are typically intended for specific readerships (e.g. L2 speakers, low-literate people, dyslexic readers), are then annotated with rules aiming at qualifying, classifying and weighting the typology of simplification operations (at lexical, syntactic and discourse-related level) encountered in the reference corpus. This is the approach followed by Caseli *et al.* (2009) for Brazilian Portuguese, Brouwers *et al.* (2014) for French and Bott and Saggion (2014) for Spanish. For other less-resourced language, e.g. Basque (Aranzabe, 2013), another approach has been pursued: it is based on the output of a readability assessment system for detecting complex sentences, which are then simplified by a large set of hand-crafted rules.

Coming to the Italian context, very few works have addressed ATS research. As we said in § 2.4, READ-IT was the first system designed with a view to text simplification and the evaluation of readability at sentence level can be considered as the first attempt to develop a semi-automatic text simplification system. The only existing fully automatic system for Italian is ERNESTA (Enhanced Readability through a Novel Event-based Simplification Tool), developed in 2013 by Barlacchi and Tonelli: this is a rule-based sentence simplification system conceived for an audience of children with poor reading skills (also called "poor comprehenders" in psycholinguistic literature), aged from 7 to 11. As this kind of readership typically struggles to select the main point and events of a story and to recover implicit information, the architecture behind ERNESTA was designed with the primary aim of highlighting *factual* events within the sentence, i.e. only the events actually happened in a story.

To this aim, a restricted set of simplification rules was implemented carrying out the following operations: *i)* sorting out pronominal and zero anaphoras, so that to make implicit content explicit; *ii)* pruning the syntactic tree, by dropping adjunct phrases and keeping only the mandatory arguments of factual verbs and *iii)* replacing past verbs by the present indicative form. An example of the output produced by ERNESTA is given below.

a) Original sentence: Ernesta stava mangiando la torta con i suoi amici.
[*Ernesta was eating the cake with her friends.*]
b) Simplified sentence: Ernesta mangia una torta.
[*Ernesta eats the cake*.]

Tested for accuracy, the system showed promising results with respect to the identification of factual events, but still presents some limits. The most problematic domain seems to be the coreference resolution, whose dedicated module in ERNESTA achieves poor performance, especially in terms of recall (i.e. 236 anaphoric elements correctly identified (i.e. 46%) out of the 515 in the test sets).

4.3 An investigation on *Italian Text Simplification*: preliminary results and perspectives

In the absence of previous research addressing ATS in Italian as a data-driven task and not by relying on a set of predefined hand-crafted rules, we designed and conducted a first exploratory study, which has tackled the following issues:

i) the design and development of a new resource to serve as a testing bed for a preliminary investigation on the process of manual text simplification and the way it is affected by the different "strategy" adopted to simplify a text, which depends, in its turn, upon who has actually simplified the text, the intended end user and the typology of texts (§4.3.1);

ii) a computational based analysis of the correlations between a set of multi-leveled linguistic features automatically extracted from the corpora and the linguistic operations retrieved in simplified corpora.

This approach has been motivated by several considerations, both theoretically and computationally driven. First, we believe that the use of parallel monolingual corpora annotated with simplification rules is not only a necessary prerequisite to detect and qualify which phenomena are involved in manual simplification before attempting to automatize the process, but it is particularly appealing from a linguistic perspective, as it allows investigating how complex linguistic phenomena are actually treated in the practice of simplification, according to different methods and target users.

The second reason moves from the consideration that typical ATS approaches, such as those described in the previous paragraph, rely on the output of a syntactic parser although the main cause of errors for an ATS system is due to erroneous parses, also when state-of-the-art parsers are used[138]. Such a concern is particular relevant in a ATS scenario, where the parsers are usually tested on domains outside of the data from which they were trained or developed on, leading to poor performance (cf. § 3.4). To give an idea of how wrong parses could affect a TS system, we should remember that the accuracy of the state-of-the art parser for Italian (i.e. DeSR parser, cf. § 2.4.1) is 87.89% in terms of Labeled Attachment Score: this corresponds to 293 erroneously parsed sentences out of the total of 376, i.e. 78% of the test sentences contain at least one parsing error.

In light of these data, the resource proposed in this study is intended to ground the development of a "semi-automatic", rather than fully automatic, TS system. Such a system, using the information extracted from the syntactic tree as only one of the features exploited to predict the rules to be applied, is expected to be more robust to syntactic parsing errors than TS system based on hand-crafted or automatically acquired rules which rely on parses transformations. Thus, it will be able to identify the areas of linguistic complexity within a sentence and suggest the authors the most appropriate simplification rule for the intended audience and domain.

4.3.1 Corpora

To tackle the first issue foreseen by our study, we started from the collection of two "parallel monolingual" corpora, called *Terence* and *Teacher*, which have to be taken as

[138] With this respect, Drndarevi´c *et al*. (2013) observed that one third of ATS errors depends on previous parsing errors and Brouwers *et al*. (2014) revealed that 89% of TS errors are due to preprocessing errors.

representative of two different text simplification strategies, i.e. a "structural" and an "intuitive" one.

These two different labels have been inspired from the definitions provided in Allen (2009), who addressed TS in the context of L2 learning. According to the author, the structural simplification relies on the use of predefined graded lists (covering both word and structural levels) or traditional readability formulas. The intuitive simplification, on the other side, is dependent on the author's teaching experience and personal judgments about the comprehension ability of learners.

In what follows we give a brief description of the two corpora and the reason way they fit these two different simplification strategies.

4.3.1.1 *Terence* corpus

The first corpus takes his name from the European project *Terence*, a three-year project devoted to designing accessible tools and resources for enhancing the comprehension of children with poor reading skills (both hearing and deaf)[139].

As reported in the project guideline, *poor comprehenders* have well-developed cognitive skills (e.g. vocabulary knowledge), though they have difficulties in deep text comprehension (e.g. inference making). One of the achievements of this project was thus the creation of suitable texts to accomplish the needs of this audience. More in detail, a corpus of 32 Italian texts, all covering short novels for children, has been manually simplified by the team of experts (psycholinguists and linguists) involved in the project, who simplified the original text at three-different, subsequent, stages:

- in the first two stages, performed by two psycholinguists, the original text was revised by focusing on the aspects of **global coherence** and **local cohesion**. The aim was to reduce the amount of inferences required by the reader and clarify the chronological order and cause/effect relationships among sentences, by introducing explicit connectives and resolving anaphoric links.

[139] More information is available at the project website: <http://www.terenceproject.eu/> [last access: 01/07/2015]

- at the last stage, the version of the original texts, already improved with respect to its global and local coherence dimensions, was further simplified in its **lexicon and syntax**. Simplification operations at this level were manually performed by a linguist and complied with a predefined guideline, specifically tailored for the audience of *poor comprehenders*. They included, among others: the replacement of unfamiliar words with more common synonyms; the elimination of idiomatic or metaphorical language; the shortening of too long and/or too complex sentences, and the substitution of unusual syntactic constructions.

Because of this controlled approach to text simplification, as well as the engagement of experts and the clearly focused target, it is possible to consider the *Terence* corpus as an appropriate instance of a "structural" text simplification strategy.

4.3.1.2 *Teacher* corpus

The second sub-corpus (i.e. *Teacher*) is composed by 24 pairs of original and simplified Italian texts, which were collected by surfing specialized educational websites providing free resources for teachers. These texts cover different textual genres, such as literature (e.g. extracts from famous novels) and descriptive texts taken from handbooks for high school on diverse subjects (e.g. history, geography).

Unlike *Terence*, text simplification was here performed independently by a different teacher, with the aim of adapting the text to the need of audience, typically L2 students with at least a B2 level in Italian. For this reason, *Teacher* exemplifies an "intuitive" simplification: while the target is usually the same (i.e. L2 learners), each text was produced by a different author and the adaptations she/he made to simplify the text affected different linguistic phenomena, without any predefined distinction or hierarchy between linguistic and textual levels.

4.3.1.3 Corpus alignment and readability evaluation

Once selected the appropriate corpora for this study, we proceeded to their alignment: for each pair of original/simplified text, the alignment was performed manually at sentence level.

More specifically, for what concerns the texts of *Terence*, we selected the two versions derived by the last two levels of simplification (i.e. local cohesion and lexicon/syntax), which were considered respectively as the original and the simplified version of our parallel corpus. This choice was motivated by the need of tackling only those textual simplification operations whose counterpart at the level of linguistic structure could be more reliably investigated by relying on linguistic features automatically extracted from the texts. On the contrary, in the case of the *Teacher* corpus, no such considerations were possible since we only had one simplified version for each original text.

Table 1 reports the alignment results.

	1:1 %	**1:2 %**	**1:3 %**	**2:1 %**	**1:0 %**	**0:1 %**
Terence	92.1	3.75	0.19	2.88	0.67	0.38
Teacher	68.32	11.45	0.76	13.74	1.15	0.0

Table 1: Percentage of sentence alignments.

As it can be noted, these data already reveal some distinctions in the way the simplification process has been carried out in the two corpora. In particular, for what concerns *Terence* (first row of Table 1), a '1:1' alignment is reported in more than 90% of cases, that is to say that the great majority of the original sentences has an exact correspondence in the simplified version of the text; 39 original sentences (3.75%) have a correspondence '1:2', thus suggesting an occurred "split" in the simplified version; only 2 original sentences underwent a three–fold split (0.19%), i.e. they correspond to three sentences in the simplified version; 15 pairs of original sentences were merged into a single one (2.88%). Finally, the percentage of unaligned sentences is 1%.

Instead, for what concerns *Teacher* (second row of Table 1), the percentage of sentences not perfectly aligned is much higher (around 25%), especially if we consider cases of sentence compression (i.e. '2:1'; '3:1'). An operation of this kind, which combines two or more independent sentences into a single one, might have been triggered by the attempt of making a text passage more cohesive, for instance when the previous sentence provides the logical background to better understand the latter. As we said in the previous paragraph, in the *Terence* corpus (cf. paragraph 4.3.1.1), the transformations aiming at improving the original texts with respect to its internal cohesion, were carried out in the first two stages of simplification, with the consequence that the original version selected for our parallel corpus

(i.e. the version resulting after these two stages) already inherited them. This is not the case of *Teacher* corpus, and the higher percentage of "merge" operations, might be interpreted as a cohesive device used by the authors to carry out the simplification.

To compare the two different simplification "strategies" with respect to the effect of the simplification process, the two corpora were also evaluated with READ-IT. More specifically, for each corpus, we calculated the value reported by the original and simplified texts on different READ–IT models (i.e. using different types of linguistic features, § 2.4) and we then calculated the Spearman's correlation between the original/simplified pairs.

As reported in Table 2, the two simplified corpora are significantly correlated with all READ–IT models. In particular, *Teacher* is especially correlated with the model using a combination of raw text and lexical features (READ–IT lexical in Table 2). This possibly follows from the "intuitive" simplification process underlying the *Teacher* corpus, which mostly concerns lexical substitution operations, given also the fact that the original texts of this corpus were much more difficult with respect to vocabulary (see the average value reported by READ-IT lexical model).

	Terence				*Teacher*			
Readability indexes	original texts	simplified texts	diff.	correlation	original texts	simplified texts	diff.	correlation
READ-IT base	0.26	0.21	0.05	**0.80***	0.45	0.19	0.26	**0.50**
READ-IT lexical	0.45	0.30	0.15	**0.65***	0.73	0.51	0.22	**0.72***
READ-IT syntax	0.20	0.16	0.04	**0.54***	0.56	0.18	0.38	**0.46**
READ-IT global	0.29	0.12	0.16	**0.77***	0.73	0.20	0.52	**0.47**

Table 2: For both *Terence* and *Teacher*, the first and the second columns report the average value obtained by the original and the simplified corpus on the corresponding READ-IT model (lower readability values indicate texts that are easier to read); the third column reports the relative difference; the fourth column shows the Spearman's correlation between the different READ-IT models and the original/simplified texts. Significant correlations ($p < 0.05$) are bolded; those with $p < 0.001$ are also marked with asterisk.

4.3.2 Simplification Annotation Scheme

After the alignment process, the original texts of both *Terence* and *Teacher* were annotated according to tagset of rules specifically designed for this study. Based on the literature on linguistic complexity and text simplification, the annotation scheme has intended to intercept, qualify and classify under different labels, a variety of transformations that a sentence possibly undergoes when it is manually simplified. As shown in Table 3, the simplification annotation scheme foresees six broad macro-categories (i.e. *split*, *merge*, *reordering*, *insert*, *delete* and *transformation*) and a specific subclass for some of them; the latter was introduced with the aim of providing a more detailed description of the linguistic level and/or typology of element affected by the rule. Such a two-leveled structure, similarly proposed by Bott and Saggion (2014) in their work on automatic text simplification for Spanish, is designed to be highly flexible and reusable, that is functional to capture both similarities and variations across paired corpora of original and abridged texts, which may result from different simplification strategies, different *genres* and different intended readerships.

Simplification Annotation Scheme	
Classes	***Sub-classes***
Split	
Merge	
Reordering	
Insert	Verb
	Subject
	Other
Delete	Verb
	Subject
	Other
Transformation	Lexical Substitution (word level)
	Lexical Substitution (phrase level)
	Anaphoric replacement
	Noun_to_Verb
	Verb_to_Noun (nominalization)
	Verbal Voice
	Verbal Features

Table 3: Simplification Annotation Scheme.

In what follows, we describe the rules covered by the annotation scheme and, for each of them, we provide an example taken from the annotated corpora[140].

Split: it is the most investigated operation in ATS, for both human- and machine-oriented applications. Typically, a split affects coordinate clauses (introduced by coordinate conjunctions, colons or semicolons), subordinate clauses (e.g., non-restrictive relative clauses), appositive and adverbial phrases. Nevertheless, we do not expect that each sentence of this kind undergoes a split, as the human expert may prefer not to detach two clauses, for instance when a subordinate clause provides the necessary background information to understand the matrix clause.

(1) O: Mamma Gorilla sembrava completamente distrutta per le cure che dava al suo vivace cuccioletto Tito, **che stava giocando vicino alle grosse sbarre di acciaio che circondavano il recinto.**
[Mummy Gorilla looked completely worn out from looking after her lively baby, Tod, **who was playing by the thick steel bars that surrounded the enclosure.**]

S: Mamma Gorilla sembrava proprio distrutta per le cure che dava al suo vivace cuccioletto Tito. *Tito stava giocando vicino alle grosse sbarre di acciaio che erano intorno alla loro area***.**
[Mummy Gorilla looked completely worn out from looking after her lively baby. *Tod was playing by the thick steel bars that surrounded the enclosure***.**]

Merge: it has to be taken as the reverse of split, i.e. the operation by which two (or more) sentences turned out to be joined into a unique one as a result of the simplification. On one side, one might expect that such a kind of transformation will be less likely adopted, as it creates semantically longer and denser sentences, which are deemed more difficult to process (Kintsh and Keenan, 1973). Yet, merging distinct sentences by providing an explicit linguistic marker can facilitate text processing (cf. § 1.4) and, to some extent (see also the discussion in § 4.4), such a rule was adopted as a simplification device in our corpora, too.

(2) O: **A causa del comportamento di Margherita, Benedetto non era troppo contento dell'organizzazione del gruppo. A Benedetto piaceva Margherita, ma non gli piaceva per niente come trattava Sandro.**

[140] In all the examples of aligned sentences taken from the corpus, "O" stands for the original and "S" for the simplified version. It is worth noting that in the majority of cases, the original sentence was affected by more than one simplification rules. Although all of them were annotated in the original sentences by marking the exact text span affected by the rule, each example focuses only on the rule under analysis: the bold string in the original sentence is the part modified by the rule, while the simplified counterpart is highlighted in italics.

[Because of Margherita's behaviour, Benedetto wasn't too happy about the group arrangement. Ben liked Margherita but he didn't really like the way she was treating Sandro.]

S: *Benedetto non era troppo contento dell'organizzazione del gruppo, perché a Benedetto non piaceva per niente il modo in cui Margherita trattava Sandro.*
[*Ben wasn't too happy about the group arrangement, because Ben didn't really like the way Maggie was treating Sean.*]

Reordering: this tag marks rearrangements of words between the original sentence and its simplified counterpart (3). Clearly, altering the position of the elements in a sentence is not an isolated event but it depends upon modifications at lexicon or syntax; e.g., replacing an object clitic pronoun (which is preverbal with finite verbs in Italian) with its full lexical antecedent[141] yields the unmarked order SVO, associated with easier comprehension and earlier acquisition (§ 1.3.3). Conversely, the editor of the simplified text may sometimes prefer a non-canonical order, when he/she believes, for instance, that it allows the reader to keep the focus stable over two or more sentences.

(3) O: Il passante gli spiegò che, per arrivare al bidone, **doveva contare ben 5 bidoni partire dal semaforo**.
The passer-by explained that, to get to the dustbin, **he had to count exactly 5 dustbins starting from the traffic light**

S: Il signore spiegò a Ugolino che *doveva contare 5 bidoni a partire dal semaforo*, per arrivare al bidone della carta.
[The man explained Little Hugh that he *had to count 5 dustbins starting from the traffic light* to get to the wastepaper dustbin.]

Insert: the process of simplification may even result in a longer sentence, because of the insertion of words or phrases that provide supportive information to the original sentence. Despite the cognitive literature suggests to reduce the inference load of a text, especially with less skilled or low-knowledge readers (Ozuru *et al*., 2009), it is difficult to predict what the author of a simple text will actually add to the sentence to make it clearer, thus making almost impossible to automatize this procedure. It can happen that the sentence is elliptical, i.e. syntactically compressed, and the difficulty depends on the ability to retrieve the missing arguments, which are then made explicit as a result of the simplification.

[141] Note that this is also a case of coreference resolution, for which a dedicated tag has been foreseen.

Our annotation scheme has introduced two more specific tags to mark insertions: one for verbs and one for subject. The latter signals the transformation of a covert subject into a lexical noun phrase (an option available in null-subject language like Italian), which might be a facilitating strategy to favour comprehension for less-skilled readers.

(4) O: Essendo da poco andata in pensione dal suo lavoro, disse che le mancavano i suoi studenti […]
[Having just retired from her job, she said that she missed her students […]]

S: Essendo da poco andata in pensione dal suo lavoro *come insegnante,* disse che le mancavano i suoi studenti […]
[Having just retired from her job *as a school teacher*, she said that she missed he students]

Delete: dropping redundant information can also be a strategy for simplifying a text. As for the *insert* rule-tag, also deletion is largely unpredictable, although we can imagine that simplified sentences would contain less adjunct phrases (e.g. adverbs or adjectives) than the authentic ones. Such occurrences have been marked with the underspecified *delete* rule (e.g. 5); two more restricted tags, *delete_verb* and *delete_subj*, have been introduced to signal, respectively, the deletion of a verb and of an overt subject (made implicit and recoverable through verb agreement morphology).

(5) O: **Sembrava veramente che** il fiume stesse per straripare.
[**It really seemed that** the river was going to burst.]

S: Il fiume stava per straripare.
[The river was going to burst.]

Transformation: this "macro-label" entails six main typologies of transformations that a sentence may undergo in order to become more comprehensible for the intended reader. Such modifications can affect the lexical, morpho-syntactic and syntactic levels of sentence representation, also giving rise to overlapping phenomena. Our annotation scheme has intended to cover the following linguistic phenomena:

- *Lexical substitution (at word level)*: when a single word is replaced by another word (or more than one), which is usually a more common synonym or a less specific term.

(6) O: Il **passante** gli spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni a partire dal semaforo.
[The **passer-by** explained that, to get to the dustbin, he had to count exactly 5 dustbins starting from the traffic light.]
S: Il *signore* spiegò a Ugolino che doveva contare 5 bidoni a partire dal semaforo, per arrivare al bidone della carta.
[The *man* explained to Little Hug that he had to count 5 dustbins starting from the traffic light, to get to the dustbin.]

Given the relevance of lexical changes in text simplification – a finding which is also confirmed by our data – previous works proposed feasible ways to automate lexical simplification, e.g. by relying on electronic resources, such as WordNet (De Belder *et al.*, 2010) or word frequency lists (Drndarevic *et al.*, 2012). However, synonyms or hypernyms replacements do not cover all the possible simplification options, since we observed that the author of the simplification might also restate the meaning of the complex word with a clause or a multi-word paraphrase, as it happened in the following case:

(7) O: Tutti **si precipitarono** verso il tendone.
[Everyone **rashed** outside the tent.]
S: Tutti si *misero a correre* verso la tenda.
[Everyone *came running* outside the tent.]

- *Lexical substitution (at phrase level)*: to capture the replacement of a whole phrase (a string made up of two or more consecutive words) with one or more words preserving the lexical meaning. In sentence (8), for instance, the rule has affected the predicative prepositional phrase, which is also a multi-word expression bearing a figurative meaning; this construct was replaced by a simple qualifying adjective, so that to avoid a potential source of difficulty for less skilled readers.

(8) O: Persino il tempo era **di buon umore.**
[Even the weather was **in a party mood**.]
S: Persino il tempo era *buono*.
[Even the weather was *good*.]

- *Anaphoric replacement*: this rule was used to annotate cases in which a referent pronoun was replaced with its full lexical antecedent (possibly a definite noun phrase or a proper noun), in order to evaluate whether this is a productive simplification strategy in our corpora. One of side, the ambiguity of pronouns might lead to misinterpret the sentence; on the other side, psycholinguistic experiments points out that the use of a lexical NP to indicate a focused-entity in the discourse can be a source itself of ambiguity at discourse level (cf. the *repeated-name penalty* effect, § 1.4.1).

(9) O: Il passante **gli** spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni […].

[The passer-by explained **him** that, to get to the dustbin, he had to count exactly 5 dustbins]

S: Il signore spiegò *a Ugolino* che doveva contare 5 bidoni a partire dal semaforo[…]

[The man explained *to Little Hug* that he had to count 5 dustbins starting from the traffic light, to get to the dustbin.]

- *Noun_to_verb*: when a nominalization or a light verb construction is replaced with a simple verb, which is easier to comprehend.

(10) O: Il giorno **della partenza**, i bambini salutarono i loro genitori durante la colazione.

[On the day **of their departure**, the children said goodbyes to their parents over breakfast.]

S: Il giorno *in cui i genitori partirono*, i bambini li salutarono durante la colazione.

[The day *that their parents left*, the children said them goodbye over breakfast.]

- *Verb_to_noun*: to mark the presence of a nominalization or a light verb construction to replace a simple verb in the original text. Although we expect that this kind of simplification operation will be less productive, since nominalizations add a degree of abstractness that can diminish the readability of texts especially for *poor readers*, it has been noted that some typologies of light verb construction are acquired very early by Italian children, particularly those introduced by the general purpose Italian verb "fare" ["to do"] (Quochi, 2007); thus, these structures might be chosen by an author instead of the equivalent, but less frequent, verb.

(11) O: La annusò dappertutto e tentò di **leccarla** [...]

He sniffed around it and attempt **to lick it** […]

S: La annusò dappertutto e le *diede una leccata* [...]
He sniffed around it and *gave it a lick* […]

- V*erbal voice*: to mark the transformation of a passive sentence into an active one or vice versa. Within both the corpora here examined, very few examples of the latter were found; this result was expected since passive sentences represent an instance of noncanonical order: they are acquired later by typically developing children (Maratsos, 1974; Bever, 1970; for Italian, (Cipriani *et al*., 1993); (Ciccarelli, 1998)) and have been reported as problematic for some atypical populations, e.g. aphasics readers (§1.3.3). Yet, the adoption of the "passivization" rule as a simplification device may be subject to the typology or genre of the texts: it can happen that the author of the simplification prefers not only to keep, but even to insert, a passive, in order to avoid more unusual syntactic constructs in Italian (such as impersonal sentences).

(12) O: Solo il papà di Luisa, "Crispino mangia cracker" era dispiaciuto, perché **era stato battuto da Tonio Battaglia**.
[Only Louise's Dad, "Cream Cracker Craig", was disappointed, because **he'd been beaten by Tod Baxter.**]
S: Solo il papà di Luisa era triste, perché *Tonio Battaglia lo aveva battuto*.
[Only Louise's Dad was sad, because *Tod Baxter had beaten him.*]

- V*erbal features:* Italian is a language with a rich inflectional paradigm and changes affecting verbal features (mood and tense) have proven useful in discriminating between easy- and difficult-to-read texts in a readability assessment task (Dell'Orletta *et al*., 2011). The easy-to-read texts examined there were also written by experts in text simplification, but their target were adults with limited cognitive skills or a low literacy level. Poor comprehenders also find it difficult to properly master verbal inflectional morphology, and the same has been noticed for other categories of atypical readers, e.g. L2 learners; thus, there is a probability that the simplification, according to the intended target, will alter the distribution of verbal features over paired sentences, as occurred in (13).

(13) O: Non capisco e **non potrei** parlare con nessuno.
[I can't understand and I **couldn't** talk to anybody.]
S: Non capisco e non *posso* parlare di queste cose con nessuno.
[I can't understand and I *can't* speak of such things to anybody.

All the rules so far described were used to annotate the two parallel corpora, so that to make it possible to analyse the productivity of each rule according to the simplification strategy adopted. These data are reported in the table below, which contains the frequency distribution of the rules within the annotated corpora.

Simplification Annotation Scheme					
Classes	***Sub-classes***	***Terence***		***Teacher***	
		%	Ab.val.	%	Ab.val.
Split		1.71	(43)	2.06	(35)
Merge		0.81	(20)	1.30	(22)
Reordering		8.65	(212)	7.89	(134)
Insert	Verb	4.92	(121)	2.53	(43)
	Subject	1.79	(44)	1.94	(33)
	Other	12.01	(295)	11.19	(290)
Delete	Verb	2.04	(50)	1.88	(32)
	Subject	0.49	(12)	0.24	(4)
	Other	19.41	(477)	23.20	(394)
Transformation	Lexical Substitution (word level)	26.50	(651)	20.73	(352)
	Lexical Substitution (phrase level)	13.39	(329)	11.60	(197)
	Anaphoric replacement	0.61	(15)	3.53	(60)
	Noun_to_Verb	1.59	(39)	0.88	(15)
	Verb_to_Noun (nominalization)	0.61	(15)	0.47	(8)
	Verbal Voice	0.53	(13)	0.77	(13)
	Verbal Features	4.92	(121)	9.78	(166)

Table 4: Percentage distribution (and absolute value) of each rule within the corpora.

A first glance on Table 4 already allows us to detect both similarities and variations across the two corpora. In particular, we can observe that the majority of rules are similarly distributed across the two corpora, thus showing that a number of simplification choices are equally chosen regardless the actual "simplifiers", i.e. a group of experts or an independent teacher. This is an interesting finding as it might suggest the existence of an "independent" simplification process shared by approaches targeting multiple audiences and based on different simplification methods.

However, there are also some exceptions, which are represented by some rules involving verbs (i.e. transformation of verbal features and insert verb) and anaphoric replacements. For what concerns the latter, it should be remembered that the *Terence* original version here adopted inherits previous sentence transformations covering, among others, anaphoric replacements (cf. § 4.3.1.1). The different distribution of rules involving verbs might instead reflect both the different simplification choices related to the "structural" and "intuitive" simplification strategies and the different textual genres included in *Teacher* and *Terence*.

4.3.3 Simplification rules and linguistic features

As this study was also intended to evaluate the reliability of using automatic readability assessment indexes for the task of (semi)-automatic text simplification, the corpora were evaluated according to a subset of multi-level linguistic features implemented in READ-IT (§ 2.4). This method allows for a more in-depth analysis of the impact and the significance of each simplification rule at sentence level.

More in detail, for both *Terence* and *Teacher*, we focused on the most frequently applied rules (i.e. Insert, Delete, Reordering, Lexical Substitution at word_level, Lexical Substitution at phrase_level) and we created as many "parallel subcorpora" as the number of rules to evaluate. For each rule, the parallel subcorpus contained the set of the original sentences to which that rule was applied and the set of the corresponding sentences resulting from the application of that rule. We then calculated the Spearman's correlation on the values of the linguistic features reported by the original and the simplified version of the sentences for each subcorpus. Table 5 (which is reported in Appendix V) illustrates the results of this analysis.

As it can be noted, all the rules are strongly correlated with the majority of linguistic features, thus suggesting that these rules do indeed have a great impact on the linguistic structure of the simplified text. Besides, the analysis also demonstrates the 'effectiveness' of such features to capture simplification operations at varying degrees of linguistic description.

Interestingly, if we examine more in-depth the significance value, we can observe a distinction between the two corpora. In particular, *Terence* reports a higher number of stronger correlations (i.e. $p < 0.001$) with respect to *Teacher*, a result that gives additional evidence to the existence of different simplification strategies, which vary according to the person (i.e. expert vs. nonexpert), textual genres and intended target.

Specifically, it seems that the teachers prefer a more vocabulary-oriented simplification approach, as testified by *a)* the highest significant correlations reported by the rules dealing with lexical replacements (i.e. *LexSub word* and *LexSub phrase*) and *b)* the fact that the majority of significant correlations at > 0.5 affects linguistic features from [1] to [19], i.e. features not dealing with the syntactic structure. Such data are likely to suggest that, independently from the simplification rule adopted, the resulting sentence has not undergone a strong modification in its grammatical structure. This is not the case of the "structural" simplification, in which all the rules significantly correlate with both lexical/morpho–syntactic features (set [1-19]) and syntactic features (set [20-35]).

On the other side, the correlation results reported by the *Delete, Lexical Substitution (word level)* and *Lexical Substitution (phrase level)* rules reveal the existence of a common approach to simplification: indeed, in the two corpora these rules are correlated with mainly the same linguistic features.

For what concerns the evaluation of the overall significance of each rule, we observe that a wide number of correlations at ≥ 0.6 occurs especially when *Split* and *LexSub word* were applied. Both these simplification operations are expected to greatly redefine the structure of the sentence: a split, e.g., not only correlates with sentence length, as it is almost expected, but it might affect the length of prepositional chains [23]. As we observed in the corpus, a potential source of a split are indeed long noun phrases linked to the main verb of the clause either as arguments or adjuncts; to simplify them the authors sometimes chose to turn them into an independent sentence, thus also adding a new verb for this sentence (see the high correlation between [23] and *InsertVerb*), as it happened in example (14).

(14) O: All'improvviso, Ernesta notò in un angolo **una strana bicicletta tutta di legno, senza pedali, e assai malconcia**!

S: All'improvviso, Ernesta notò in un angolo una strana bicicletta. ***Era*** *tutta di legno, senza pedali, e molto rovinata.*

4.4 Discussion

The aim of this work was to design a flexible, theoretically informed and machine-usable annotation scheme capable of detecting and classifying under distinct labels a wide range of linguistic phenomena involved in text simplification as a previous step to automatize this process.

The preliminary results here obtained seem quite promising in this sense: in particular, the analysis of the correlations – from which it is was possible to isolate those patterns of linguistic features that are more affected by specific simplification rules – makes it possible to conceive a text simplification system that should be able not only to identify the areas of complexity within a sentence, but also to suggest a possible rewriting informed by the patterns of feature distribution learnt from real simplified texts.

However, the availability of more and larger parallel corpora is a necessary condition to allow for a more granular classification of the simplification phenomena, which is an aspect that deserves a careful attention.

With this respect, let's focus again on the frequency distribution of the simplification rules in the corpora (Table 4): as we can see, one of the most evident data is the large exploitation of operations dealing with lexical substitutions, particularly at word level. A thorough analysis of these operations is clearly required not only to refine the phenomena currently annotated under this macro-class (especially when they affect the transformation of a word into a sequence of words) but also to investigate whether there is a correspondence between the concept of linguistic complexity/simplicity as it emerges from the literature and the way "complex" structures are treated in the practice of simplification, given the text at hand.

Let's consider, e.g., sentence (14), in which the bolded string has been currently annotated as an instance of the rule defined as "Lexical Substitution (at phrase level)"; yet, it has to be noted that the substituted string (containing an abstract noun) has been paraphrased with a subordinate clause, and even a typology of complex clause, i.e. an indirect relative clause. Similarly, in (15), the level of lexical complexity has been reduced by eliminating the technical term (i.e. calamine), probably unknown to children, which has been explained by means of a gloss; yet, this choice led to the insertion of a final adverbial clause modifying the NP, which might increase complexity at syntactic level.

(14) O: La gente poteva ridurre i propri rifiuti **comprando solo il necessario** […]
[People could reduce their rubbish by **buying only the necessary [things]**].

S: La gente poteva ridurre i propri rifiuti **non comprando cose di cui non aveva bisogno**.
[People could reduce their rubbish by **not buying things they didn't need.**]

(15) O: Pensò che forse la lozione **alla calamina** l'aveva rimpicciolita durante la notte.
[She thought that maybe **the calamine lotion** had shrunk her overnight.]

S: Pensò che forse la lozione **per calmare il prurito** l'aveva fatta diventare più piccola durante la notte.
[She thought that maybe the lotion **to stop her itching** had made her smaller in the night.]

For this typology of sentences, and especially those exemplified by (14), where the simplification also gives rise to the "tricky" double negative construction, it should be worth validating the effect of the adopted simplification device by means of empirical tests with the intended readers.

Interestingly, examples like (14) and (15) also highlight how the effect of the simplification can result in a longer sentence, thus giving additional confirmation that the direct relationship between sentence complexity and sentence length is not straightforward. This is especially true when we consider the output resulting from the application of the "merge" simplification rule.

With this respect, a qualitative analysis of these sentences seems to suggest that, in this case too, a subcategorization of the "merge" annotation tag should be possible. We observed, indeed, instances in which the simplified sentence stemming from two distinct sentences in the original text does not impact at the level of text processing and the inferential load required (e.g. (16), (17)).

(16) O: **Confidò a Luisa, la sua nuova amica, che erano secoli che desiderava farsi tagliare i capelli. Erano così caldi e pesanti**!
[She confided to Louise, her new friend, that she had been wanting to cut her hair cut off for ages. It was so hot and heavy!]

S: *Disse a Luisa, che ora era la sua nuova amica, che aveva sempre desiderato tagliarsi i capelli: erano così caldi e pesanti!*
[She told Louise, who was now her new friend, that she had wanted to cut her hair off: it was so hot and heavy!]

(17) O: **Clara pensò che fosse uno dei cigni. Ma poi si rese conto che stava urlando!**
[Clara thought it was one of the swans. But then she realised it was shouting!]

S: *In un primo momento, Clara pensò che fosse uno dei cigni ma poi si rese conto che stava urlando!*
[At first, Clara thought it was one of the swans, but then she heard it shouting!]

In other cases, instead, the modified sentence exhibits an explicit connective marker that clarifies the logical link (especially causal and temporal) between the first and the second propositional unit, thus resulting in a simplification that, despite the increased sentence length, makes the passage more coherent by reducing the inference generation process, which can be hard for less-skilled readers (e.g. (18), (19)).

(18) O: **Ida si sentì stupida mentre sprofondava di nuovo nel sonno. "Devo aver sognato", si disse.**
[Ida felt silly as she drifted back off to sleep. "I must have been dreaming", she told herself.]

S: *Ida si sentì stupida mentre tornava a dormire perché pensava di aver sognato*.
[Ida felt silly as she went back to sleep because she thought she must have been dreaming.]

(19) O: **Pensò di volare sopra i tetti della città. Puntò ancora una volta i piedi a terra e si spinse in avanti a tutta velocità, con la testa bassa, ben incassata fra le spalle**.
[She thought of flying over the roofs of the town. She pushed her feet on the ground once more and she went forwards at full speed, with her head down, steady between her shoulders.]

S: *Mentre pensava di volare sopra i tetti della città, spinse i piedi a terra, e si mise a correre veloce con la bicicletta*.
[*While* she was thinking of flying over the roofs of the town, she pushed her feet on the ground and started riding quickly on the bicycle.]

Chapter 5

Conclusion and future perspectives

This thesis has been concerned with two general issues: what linguistic complexity is for a human reader and how it is possible to operationalize it by means of language technologies, in a way that is capable of giving theoretical awareness to human-oriented applications for improving text accessibility.

The attempt of unifying the cognitive perspective and the computational linguistics perspective to the study of linguistic complexity has represented the main challenge of the whole work and provided the leitmotiv around which the reference literature has been reviewed in the first part. More specifically, Chapter 1 has addressed linguistic complexity in terms of *processing* difficulty, thus highlighting a wealth of properties pertaining to lexical, syntactic and textual objects, which modern linguistic research in the cognitive framework has identified as involved in the process of human language comprehension. In Chapter 2, such a set of formal properties has been inspected by relying on the "tools" (in terms of resources, algorithms, methods) drawn from computational linguistics, and specifically from the viewpoint of research in automatic readability assessment of written texts. A selective review of recent approaches to this field has allowed us to endorse the possibility of modeling a large set of linguistic complexity predictors, which are effective not only in terms of their computational treatability but also in light of their explanatory power with respect to cognitive awareness.

Once established the adequacy of the methodological "apparatus" to support automatic text difficulty analysis, we have proceeded with the elaboration of the leading research questions, first outlined in the abstract of the thesis and repeated below, i.e.:

- Which features of a text embody a general, i.e. valid "across textual genres", notion of linguistic complexity?
- Which features of a text embody a "genre-specific" notion of linguistic complexity, such as the one characterizing the domain of bureaucratic language?

- Is this twofold typology of features already handled by current readability assessment indexes? And, if the answer is no, how can an automatic system succeed in learning the difference?
- In what way linguistic complexity features for readability assessment can make it possible the automation of related and, more specific applicative tasks, such as text simplification?

This is what has been done in Chapter 3 and Chapter 4, which constitute the most original contributions of this study, and where such broad questions have been concretely put at work in two applicative tasks.

The first one has investigated the potential impact of genre-specific features automatically extracted from texts on general-purpose readability assessment tools, by conducting a linguistic profiling investigation of a "quasi-parallel" corpus of Italian texts belonging to the bureaucratic language variety. Despite the extensive literature devoted to simplifying bureaucratic language, we have seen how the controversial status of bureaucratic language as a "special language" makes it not trivial to establish which properties, although deemed proxies of linguistic complexity in ordinary texts, are yet required to accomplish the formal requirements of these texts. By comparing how these properties – modelled in terms of features automatically extracted from the multi-leved output of linguistic annotation – give rise to both different and similar patterns of distributions within the corpus under examination, and with respect to a monitor corpus, it was possible to distinguish "genre-complexity" markers from the stylistic signals of *bureaucratese*. In this regard, although data were collected from a small corpus, we noted that several of the emerged tendencies resembled what already highlighted by traditional analytical approaches, thus makes it possible to corroborate the effectiveness of this methodology as a research tool for large-scale studies into language variation across genres. This is an important result, which is expected to be further bolstered by adapting the tools of linguistic analysis to the characteristics of the domain at hand.

Besides, for what concerns the potential applicative outcome of this study, the results of the linguistic profiling wish to serve as the starting point to ground the specialization of a readability index tailored to the bureaucratic genre. Such a tool, linguistically informed, could contribute to enhance the communication process between citizens and institutions, which is a concern of major importance in modern information society, particularly with respect to specific groups of readers.

And especially with regard to the issue of text accessibility – where language technologies can play a fundamental role – linguistic research in the domain of cognitive science, language acquisition and neurolinguistics highlights the role of fine-grained factors of linguistic complexity, such as the discussed locality effects in long distance syntactic dependencies, which are particularly taxing and can obstacle a deep understanding of texts by language-impaired speakers, e.g. agrammatic aphasics. A further advancement in automatic readability assessment research would thus require to personalize models and to evaluate whether actual metrics, even when tailored for genre, are also robust enough to capture processing difficulties at higher-level of comprehension for specific users. For such a goal to be fulfilled, it seems quite necessary to train a system on gold corpora, whose score of complexity has been assessed both by experts and by means of comprehension tests addressed to the intended target.

A similar concern holds for research in automatic text simplification, a NLP technology that is even at an earlier stage of growth compared to computational readability assessment. In Chapter 4, we have described preliminary work, which has led to the development of a first resource for the Italian language, where a classification of multi-leveled simplification operations has been established starting from the analysis of two parallel monolingual corpora, intended for different targets. Despite some refinements of the annotation scheme should be considered, according to the qualitative observations outlined in § 4.4, we believe that this study opens up interesting research perspectives, both at theoretical and applicative level. One of these would be to evaluate whether the proposed annotation scheme works well with other samples of original/simplified corpora, representative of different textual typologies. With this respect, a quite natural step unifying the different tasks tackled in this thesis would be to test the suitability of the annotation scheme on the bureaucratic corpus described in Chapter 3. The latter was there defined as a "quasi-parallel" corpus, since it contained significant misalignments at sentence level due to the multiple stages of adaptations the original text underwent; this is way we did not rely on it as the initial testing bed for the scheme. However, it should be possible to extract a sub-corpus containing only paired sentences and annotate it with the rules foreseen in the scheme, so that to verify both their productivity and the capability to intercept simplification phenomena triggered by the presence of *bureaucratese* features.

Along with *cross-domain* text simplification, also *cross-language* text simplification is an attractive avenue to explore. Parallel monolingual corpora annotated with rules comparable to those proposed in our scheme exist today for e.g. Spanish, French, Basque (cf. § 4.2) and

could be quite easily obtained for English too, by using the already available aligned resources, (e.g. the English Wikipedia/ Simple English Wikipedia, first of all). This makes it possible to investigate new research questions, such as: are there some "universal" tendencies in the way texts come up to be simplified, possibly similar to those characterizing nonstandard simplified registers, e.g. the child-directed speech (i.e. the so-called "motherese")? Despite the different realizations in languages, are there comparable phenomena, or a "hierarchy" of phenomena, that trigger a particular simplification operation? If this is the case, what kind of features allows for capturing the effects of *cross-language* simplification, when different formalisms have been used to create an automatically annotated version of the corpus (e.g. dependency vs. constituency parsers)?

It is clear that the scope of these questions goes beyond the specific task of text simplification; instead, it can contribute to bring new perspectives from which exploring and enriching our understanding of the concept of linguistic complexity.

APPENDIX I

The morpho-syntactic tagset here reported, as well as the dependency tagset (Appendix II), were jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) and the University of Pisa in the framework of the TANL (*Text Analytics and Natural Language processing*) project and they were used for the annotation of the ISST-TANL dependency annotated corpus, which originates as a revision of the ISST-CoNLL corpus.

ISST-TANL morpho-syntactic tagset

The ISST-TANL part-of-speech tags are based on the ILC/PAROLE tagset and are conformant to the EAGLES international standard. The table below documents the 14 coarse-grained pos tags (column 1) and the 37 fine-grained tags (column 2) used for ISST-TANL annotation.

Coarse-grained tag	**Fine-grained tag**	**Description**	**Examples**	**Contexts of use**
A	A	Adjective	*bello, buono, pauroso, ottimo*	una **bella** passeggiata un **ottimo** attaccante una persona **paurosa**
	AP	possessive adjective	*mio, tuo, nostro, loro*	a **mio** parere il **tuo** libro
B	B	Adverb	*bene, fortemente, malissimo, domani*	arrivo **domani** sto **bene**
	BN	negation adverb	*non*	**non** sto bene
C	CC	coordinative conjunction	*e, o, ma, ovvero*	i libri **e** i quaderni vengo **ma** non rimango
	CS	subordinative conjunction	*mentre, quando*	**quando** ho finito vengo **mentre** scrivevo ho finito l'inchiostro
D	DE	exclamative determiner	*che, quale, quanto*	**che** disastro! **quale** catastrofe!
	DI	indefinite determiner	*alcuno, certo, tale, parecchio, qualsiasi*	**alcune** telefonate **parecchi** giornali **qualsiasi** persona
	DQ	interrogative determiner	*che, quale, quanto*	**che** cosa **quanta** strada **quale** formazione
	DR	relative determiner	*cui, quale*	i **cui** libri
	DD	demonstrative determiner	*questo, codesto, quello*	**questo** denaro **quella** famiglia
E	E	Preposition	*di, a, da, in, su, attraverso, verso, prima_di*	**a** casa **del** poeta **prima_di** giorno **verso** sera
	EA	articulated preposition	*del, alla, dei, nelle*	**nella** casa il prezzo **del** pane
F	FB	"balanced" punctuation	( ) " " ' ' - -	il gatto – che conoscete –

Coarse-grained tag	Fine-grained tag	Description	Examples	Contexts of use
	FC	clause boundary punctuation	*, ;*	ha detto **:** Vieni!
	FF	comma, hyphen	*,*	mele**,** pere e banane due-trecento persone
	FS	sentence boundary punctuation	*. ? !*	mele, pere e banane**.** cosa vuoi**?**
I	I	Interjection	*ahimè, beh, ecco, grazie*	**Beh**, che vuoi?
N	N	cardinal number	*uno, due, cento, mille, 28, 2000*	**due** partite **28** anni
	NO	ordinal number	*primo, secondo, centesimo*	**secondo** posto
P	PD	demonstrative pronoun	*questo, quello, costui*	**quello** di Roma **costui** uccide
	PE	personal pronoun	*egli, lui, esso* *noialtri, voialtri, essi* *io, me, tu, te*	**io** parto **lo** mangio
	PI	indefinite pronoun	*chiunque, ognuno, molto*	**chiunque** venga i diritti di **ognuno**
	PP	possessive pronoun	*mio, tuo, suo, loro, proprio*	il **mio** è qui più bella della **loro**
	PQ	interrogative pronoun	*che, chi, quanto*	non so **chi** parta **quanto** costa? **che** ha fatto ieri?
	PR	relative pronoun	*che, cui, quale*	**ciò** che dice il **quale** afferma a **cui** parlo
	PC	clitic pronoun	*ci, vi, mi, ti, la, le*	**lo** vidi **li** ho sentiti aver**la** **le** dissero, **le** videro **mi** dicono **ci** sposiamo **vi** credo **si** sente, **si** sentono **ci** vado spesso
R	RD	determinative article	*il, lo, la, i, gli, le*	**il** libro **i** gatti
	RI	indeterminative article	*uno, un, una*	**un** amico **una** bambina
S	S	common noun	*amico, insegnante, verità*	l'**amico** la **verità**
	SA	abbreviation	*ndr, a.C., d.o.c., km*	30 **km** sesto secolo **a.C.**
	SP	proper noun	*Monica, Pisa, Fiat, Sardegna*	**Monica** scrive
T	T	predeterminer	*tutto, entrambi, ambedue*	**tutte** le notizie **ambedue** le idee
V	VA	auxiliary verb	*avere, essere, venire*	il peggio **è** passato **ho** scritto una lettera

Coarse-grained tag	Fine-grained tag	Description	Examples	Contexts of use
				viene fatto domani
	VM	modal verb	*volere, potere, dovere, solere*	non **posso** venire **vuole** il libro
	V	main verb	*mangio, avere, passato, camminando*	il peggio è **passato** ho **scritto** una lettera **vengo** domain
X	X	residual class	it includes formulae, unclassified words, alphabetic symbols and the like	distanziare di **43''** mi **piacce**

APPENDIX II

ISST-TANL Syntactic dependency annotation tagset

Tag	Relation Type	Description	Examples
Arg	Argument	Relation between a verbal or nominal head and a non-subject clausal argument.	Il 63% dei francesi ha **imposto** al presidente **di** rinunciare alla sua bomba È giunto il **momento di** creare un'area denuclearizzata Le autorità hanno **annunciato che** il blitz è concluso La **decisione di** continuare... **escludendo che** il militare volesse veramente mettere in pericolo... si sono **rifiutati di** fornire informazione
Aux	auxiliary	Relation between a verb and its auxiliary.	Il corazziere **è stato** individuato Il corazziere è **stato individuato** **Ha dichiarato** di **aver pagato** i terroristi
Clit	clitic	Relation between a clitic pronoun and a verbal head used in pronominal form.	La sedia **si** è **rotta** Non **ci rendiamo** conto **Si tratta** della scoperta
comp	complement	Relation between a head and a prepositional complement, whether a modifier or a subcategorized argument.	Fu **assassinata da** un pazzo E' più **interessante del** libro **Oggi come** allora **Più di** quattrocento esemplari **Osteggiata dal** governo di Berna **Grande quanto** mezza Italia
comp_ind	indirect complement/object	Denotes the affected participant of an event.	Ho **dato** il libro **a** lui I carabinieri **gli** hanno **recapitato** il decreto
comp_loc	locative complement	Expresses either a location or a direction of movement of an action.	Si **trovava in** un parco Era **uscito di** casa alle 10
comp_temp	temporal complement	Denotes a temporal relation with a verbal head.	**Nel** 1985 è stata **uccisa** un'antropologa L'allarme è **scattato** la scorsa **settimana**
Con	copulative conjunction	Relation between a copulative conjunction in coordinate structures and the first conjunct (which becomes the head of the whole coordinate structure).	Una ragazza **violentata e** sequestrata da due slavi **Gabriella e** Paolo sono partiti Hanno **riarmato,** addestrato e preparato l'esercito Hanno riarmato, **addestrato e** preparato l'esercito **Scontri,** assalti e centinaia di feriti Scontri, **assalti e** centinaia di feriti
concat	Concatenation	Relation between tokens forming complex word forms (e.g. complex proper nouns, multi-word expressions and the like).	Il segretario di **De Michelis** L'enciclica "**Mulieris dignitatem**" La **International Public** Sport La International **Public Sport**
conj	conjunct linked by a copulative	Relation between the conjuncts after the first to the first one, which is the	Una ragazza **violentata** e **sequestrata** da due slavi **Gabriella** e **Paolo** sono partiti

	conjunction (con)	head of the whole coordinate structure. conj is used in association with coordinating copulative conjunctions.	Hanno **riarmato**, **addestrato** e preparato l'esercito Hanno riarmato, **addestrato** e **preparato** l'esercito **Scontri**, **assalti** e centinaia di feriti Scontri, assalti **e centinaia** di feriti
Det	determiner	Relation between a nominal head and its determiner.	**Una sala** ha dovuto essere sgomberata Rilevata **la presenza** di gas
Dis	disjunctive conjunction	Relation between a disjunctive conjunction in coordinate structures and the first conjunct which is taken to be the head of the whole coordinate structure.	Cassonetti dell'immondizia **rovesciati o** incendiati Partecipa **a** manifestazioni politiche **o** a dibattiti
Disj	conjunct in a disjunctive compound linked by a disjunctive conjunction (dis)	Relation between the (second, third, …) conjuncts to the first conjunct which is taken as the head of the whole coordinate structure. disj is used in association with coordinating disjunctive conjunctions.	Cassonetti dell'immondizia **rovesciati** o **incendiati** Partecipa **a** manifestazioni politiche o **a** dibattiti
mod	Modifier	Relation between a head and its adjectival, adverbial, and clausal modifier. For example, noun+adjective, adverb+verb, and phrasal modifiers. Also noun+noun appositive constituents.	I colori **sono sempre** gli stessi **Colori intensi** Trionfo di Didoni nei **20 km** di marcia **Cesare** l'**Imperatore** **Per** arrivare in tempo, sono **partito** molto presto **Quando** la campanella suona, i bambini **escono** da scuola
mod_loc	locative modifier	Relation between a head and its adjectival, adverbial, and clausal modifier that expresses either a static or directional location.	Non **so dove** Tutto **cominciò** proprio **lì** Avrei voluto **fermar**mi **qui** più a lungo
mod_rel	relative modifier	Relation between the verbal head of a relative clause and its nominal head in the higher clause. The mod_rel relation is also used in case of free relatives, linking the verbal head of the free relative to the chi pronoun (which in turn is directly linked to its governor)	**Box** che è stato **trovato** nel pomeriggio Quell'**ordine** che i due Stranamore pentiti avevano **imposto** per cinquant'anni Non è mai stato accertato **chi volle** la sua morte
mod_temp	temporal modifier	A temporal relation between a head and its adjectival, adverbial, and clausal modifier.	**Ieri** hanno **dormito** all'aperto Scoperto 75 **anni fa** Non **superano mai** gli 8 milioni
modal	modal verb	Relation between a verbal head and a modal verb.	Una sala ha **dovuto** essere **sgomberata** Avrebbe **potuto ripeter**si

Neg	Negative	Negative modifier (“no” or “non”)	A volte **non dormo**
Obj	direct object	Relation between a verbal head and its direct object (always non-clausal).	**Hanno** un **modo** di ragionare rozzo **Centellinando** le **informazioni** È giunto il momento di **creare** un'**area** denuclearizzata **Rilevata** la **presenza** di gas
pred	predicative complement	Relation between a head and a predicative complement, be it subject or object predicative.	L'incontro è **stato fatale** Questo **è** il **messaggio** finale
pred_loc	locative predicate	Expresses a spatial property of the subject, after a linking verb.	Il presidente non **era in** casa
pred_temp	temporal predicate	Expresses a temporal property of the subject, after a linking verb.	La riunione **è alle** 5
prep	Preposition	Relation between a prepositional head and its complement, whether clausal or non-clausal.	Un contributo **alla lotta** contro la criminalità Un contributo alla lotta **contro** la **criminalità** Prima **di partire** ho telefonato
punc	Punctuation	Relation between a word token and a punctuation mark.	**Teatro** della tragedia **,** ...
ROOT	sentence root	Head of sentence.	**Desidero** dormire Note that only the dependent is shown, since the head is a fictitious root node
Sub	subordinate clause	Relation between a subordinative conjunction and the verbal head of its clausal complement.	Ha detto **che** non **intendeva** fare nulla Le autorità hanno annunciato **che** il blitz è **concluso** Venne ucciso **mentre cercava** di difendere la ragazza
subj	subject	Relation between an active verb and its subject. It is also used to mark clausal subjects. When the subject is not explicit, as it occurs in pro-drop languages like Italian, the subject relation is not present: the morpho-syntactic features of the subject, can be induced from the inflectional features of the verb.	il **testimone** ha **parlato** subito le **vittime seguivano** gli aiuti
subj_pass	passive subject	Relation between a passive verb and its subject.	I **missionari** erano stati **rapiti** la mattina presto Circa 83.000 **franchi** furono **spesi**

APPENDIX III

This table contains the percentage distribution values of the major ('coarse-grained') morpho-syntactic categories for the corpora described in Chapter 3 (§ 3.6.1.1). In particular, the last column illustrates the results of the bureaucratic corpus, with respect to the simplified (left side) and the original (right side) version.

PoS Distribution	***Journalism***		***Educational***		***Literature***		***Scientific Prose***		***Legal Language***		***Bureaucracy***	
	Due Par	Rep	Edu_ Child	Edu_ Adult	Narr Child	Narr Adult	Wiki	Scient Art	Norm_ acts	It_ Const	Bur Simp	Bur orig
Adjectives	5.92	6.40	6.61	8.81	5.93	6.38	8.71	8.99	8.16	8.40	5.72	5.98
	6.16		7.76		6.15		8.85		8.28		5.85	
Adverbs	3.52	4.83	5.72	5.86	6.43	5.38	4.15	3.81	1.43	2.29	1.91	2.14
	4.18		5.79		5.90		3.98		1.86		2.03	
Conjunctions	3.69	3.61	4.37	5.01	4.83	4.36	3.75	3.44	4.13	5.29	3.09	2.75
	3.65		4.69		4.60		3.60		4.71		2.92	
Determiners	1.65	0.84	1.03	1.19	1.21	0.90	1.11	1.30	0.50	0.81	0.69	1.02
	1.24		1.11		1.06		1.21		0.65		0.85	
Prepositions	15.28	16.41	13.89	15.25	12.09	12.34	16.46	17.63	20.64	18.63	18.07	20.42
	15.85		14.57		12.21		17.05		19.64		19.25	
Punctuations	10.96	12.92	12.62	11.63	14.05	15.24	12.95	11.51	10.97	8.83	11.93	11.07
	11.94		12.13		14.65		12.23		9.90		11.50	
Interjections	0.01	0.04	0.08	0.03	0.20	0.08	0.04	0.06	0.00	0.00	0.01	0.00
	0.03		0.06		0.14		0.05		0.00		0.00	
Numbers	2.73	2.39	1.02	0.80	1.10	1.65	1.74	2.71	6.00	2.77	6.57	5.22
	2.56		0.91		1.37		4.44		4.38		5.90	
Pronouns	2.32	3.76	5.11	5.61	6.88	6.13	3.18	2.88	2.09	2.40	2.74	3.18
	3.04		5.36		6.51		3.03		2.25		2.96	
Articles	10.39	8.31	9.48	8.57	8.35	8.07	8.29	7.17	6.63	8.55	6.42	5.94
	9.35		9.03		8.21		7.73		7.59		6.18	
Nouns	29.30	27.19	23.17	22.99	21.96	24.08	28.46	28.41	30.27	30.16	30.52	29.82
	28.24		23.08		23.02		28.44		30.21		30.17	
Predeterminers	0.25	0.08	0.11	0.17	0.16	0.10	0.13	0.08	0.08	0.18	0.06	0.09
	0.17		0.14		0.13		0.11		0.13		0.08	
Verbs	13.66	12.89	15.05	12.67	15.83	14.96	10.65	10.60	8.59	11.50	11.05	11.12
	13.28		13.86		15.39		10.62		10.04		11.09	
Residuals	0.01	0.00	0.00	0.01	0.00	0.03	0.02	0.23	0.43	0.00	0.09	0.11
	0.00		0.00		0.01		0.13		0.22		0.10	

Table 5: Percentage distribution of major morpho-syntactic categories across all the corpora.

APPENDIX IV

This table displays the percentage distribution values of the syntactic dependency relationships for the corpora described in Chapter 3. In particular, the last column reports the results of the bureaucratic corpus, with respect to both the simplified (left side) and the original (right side) version.

Syntactic Depend. rel	***Journalism***		***Educational***		***Literature***		***ScientificProse***		***Legal lang***		***Bureaucracy***	
	Due Par	*Rep*	*Edu_ Child*	*Edu_ Adult*	*Narr_ child*	*Narr_ Adult*	*Wiki*	*Scient Art*	*It_ Const*	*Norm_ acts*	*Bur_ Simp*	*Bur_ Orig*
ROOT	6.00	5.31	6.10	4.32	8.67	10.34	4.85	4.71	6.63	5.90	6.01	4.63
	5.66		5.21		9.50		4.71		6.26		5.32	
Arg	1.57	1.81	1.84	1.84	2.30	2.11	1.03	1.28	0.70	1.03	1.53	1.81
	1.69		1.84		2.20		1.16		0.87		1.67	
Aux	2.19	2.14	1.78	1.10	1.22	1.62	1.03	1.23	1.73	1.02	1.29	1.19
	2.16		1.44		1.42		1.13		1.37		1.25	
Clit	0.48	0.86	1.19	1.19	1.52	1.20	0.81	0.73	0.47	0.27	0.45	1.53
	0.67		1.19		1.36		0.77		0.37		0.99	
Comp	11.26	13.35	11.18	12.55	9.84	9.89	14.11	15.45	15.65	17.18	14.95	16.63
	12.30		11.87		9.86		14.78		16.42		15.79	
Comp_ind	0.09	0.20	0.26	0.15	0.52	0.49	0.05	0.06	0.12	0.03	0.29	0.14
	0.14		0.20		0.51		0.06		0.08		0.22	
Comp_loc	1.16	0.73	0.58	0.36	0.42	0.49	0.46	0.32	0.32	0.34	0.99	0.96
	0.94		0.47		0.45		0.39		0.33		0.97	
Comp_temp	0.84	0.45	0.38	0.19	0.18	0.19	0.30	0.15	0.35	0.33	0.64	0.81
	0.64		0.29		0.19		0.22		0.34		0.73	
Comp_con	4.07	3.70	5.22	5.30	4.97	4.47	4.08	3.50	4.79	4.21	2.56	2.51
	3.89		5.26		4.72		3.79		4.50		2.54	
Comp_concat	0.06	0.12	0.01	0.06	0.09	0.08	0.26	0.24	0.03	0.04	0.06	0.12
	0.09		0.03		0.08		0.25		0.04		0.09	
Comp_conj	4.07	3.30	4.40	4.62	3.88	3.67	3.87	3.34	4.35	3.73	2.66	2.70
	3.68		4.51		3.77		3.60		4.04		2.78	
Det	10.37	8.30	9.46	8.56	8.32	8.05	8.29	7.14	8.48	6.61	6.42	5.94
	9.34		9.01		8.19		7.72		7.54		6.18	
Dis	0.28	0.11	0.20	0.31	0.16	0.18	0.55	0.29	0.87	0.59	0.31	0.22
	0.20		0.25		0.17		0.42		0.73		0.26	
Disj	0.23	0.07	0.17	0.22	0.11	0.11	0.35	0.18	0.75	0.44	0.26	0.16
	0.15		0.19		0.11		0.26		0.60		0.21	
Mod	16.55	17.74	14.94	17.48	15.98	15.70	19.28	21.20	15.72	20.35	21.31	19.80
	17.15		16.21		15.84		20.24		18.04		20.56	
Mod_loc	0.04	0.06	0.16	0.09	0.14	0.12	0.05	0.06	0.00	0.01	0.01	0.01
	0.05		0.12		0.13		0.05		0.00		0.01	
Mod_rel	1.31	1.35	1.57	1.81	1.45	1.43	1.44	1.23	0.87	0.84	0.68	0.75
	1.33		1.69		1.44		1.33		0.85		0.72	
Modal	0.73	0.45	0.45	0.64	0.55	0.49	0.61	0.55	0.99	0.52	0.79	0.66
	0.59		0.54		0.52		0.58		0.75		0.72	
Neg	0.34	0.69	0.59	0.93	1.01	1.12	0.39	0.40	0.93	0.32	0.48	0.49
	0.51		0.76		1.07		0.40		0.63		0.48	
Obj	4.86	3.52	3.81	3.69	4.65	4.25	2.78	2.86	3.46	2.41	3.11	2.39
	4.19		3.75		4.45		2.82		2.93		2.75	
Pred	1.63	1.47	2.56	2.00	1.87	1.88	1.93	1.37	0.98	0.52	0.64	0.81
	1.55		2.28		1.87		1.65		0.75		0.73	
Prep	15.28	16.39	13.84	15.24	11.99	12.17	16.45	17.53	18.65	20.50	18.05	20.41
	15.84		14.54		12.08		16.99		19.57		19.23	
Punc	9.58	11.6	10.32	9.88	11.90	13.03	11.43	10.34	7.72	9.54	11.10	10.30
	10.63		10.10		12.47		10.88		8.63		10.70	

Syntactic Depend. Rel	***Journalism***		***Educational***		***Literature***		***Scientific Prose***		***Legal lang***		***Bureaucracy***	
	Due Par	*Rep*	*Edu_ Child*	*Edu_ Adult*	*Narr_ child*	*Narr_ Adult*	*Wiki*	*Scient Art*	*It_ Const*	*Norm_ acts*	*Bur_ Simp*	*Bur_ orig*
Sub	0.63	0.87	0.97	1.03	1.34	1.23	0.57	0.65	0.67	0.39	0.94	0.78
	0.75		1.00		1.29		0.61		0.53		0.86	
Subj	5.45	4.21	5.46	4.58	5.31	4.83	4.02	3.38	3.45	2.12	2.60	2.39
	4.83		5.02		5.07		3.70		2.78		2.50	
Subj_pass	0.20	0.31	0.54	0.25	0.20	0.16	0.56	0.53	1.00	0.53	0.35	0.51
	0.26		0.40		0.18		0.54		0.77		0.43	

Table 11: Percentage distribution of the main syntactic dependency relationships across all the corpora.

APPENDIX V

This table reports the Spearman's correlation between the most frequent simplification rules foreseen by the simplification annotation scheme and a subset of linguistic features (Chapter 4, § 4.3.3). Significant correlations ($p < 0.05$) are bolded; those with $p < 0.001$ are also marked with asterisk (*). For each column, the left value refers to the *Terence* corpus, the right value to the *Teacher* corpus.

Features	**Insert**		**Delete**		**Reordering**		**LexSub word**		**LexSub phrase**		**Split**		**Insert Verb**	
[1] Sentence length	**.796***	**.342**	**.772***	**.345***	**.820***	**.451***	**.818***	**.463***	**.787***	**.433***	**.799***	**.501**	**.714***	**.573***
[2] Word length	**.595***	**.431***	**.593***	**.518***	**.627***	**.637***	**.636***	**.559***	**.512***	**.449***	**.700***	**.581**	**.612***	**.375**
[3] Word types in BIV	**.663***	**.315**	**.707***	**.382***	**.699***	**.456***	**.735***	**.580***	**.654***	**.472***	**.630***	**.865***	**.690***	**.413**
[4] Lexical density	**.639***	**.246**	**.685***	**.416***	**.704***	**.410***	**.757***	**.400***	**.617***	**.402***	**.646***	**.696***	**.566***	.082
[5] Adjective	**.693***	**.450***	**.689***	**.406***	**.752***	**.564***	**.724***	**.585***	**.726***	**.527***	**.779***	**.662**	**.787***	.245
[6] Adverbs	**.546***	**.324**	**.652***	**.424***	**.667***	**.311**	**.729***	**.445***	**.581***	**.245**	**.670***	.292	**.716***	**.351**
[7] Coord Conj	**.609***	**.345**	**.707***	**.454***	**.735***	**.588***	**.765***	**.554***	**.746***	**.494***	**.474**	**.662**	**.667***	.306
[8] Subord Conj	**.510***	**.532***	**.611***	**.478***	**.564***	**.606***	**.700***	**.483***	**.716***	**.414***	**.726***	**.554**	**.641***	**.441**
[9] Preposition	**.687***	**.492***	**.678***	**.404***	**.690***	**.354**	**.794***	**.498***	**.680***	**.447***	**.688***	**.491**	**.743***	**.480**
[10] Pronoun	**.619***	.179	**.629***	**.277**	**.550***	**.304**	**.716***	**.317***	**.594***	**.338***	**.552***	**.578**	**.368***	-.030
[11] Noun	**.707***	**.566***	**.702***	**.586***	**.708***	**.474***	**.761***	**.601***	**.721***	**.548***	**.666***	**.544**	**.728***	**.490**
[12] Verb	**.703***	**.401***	**.634***	**.464***	**.655***	**.435***	**.722***	**.506***	**.653***	**.468***	**.743***	**.679**	**.656***	.268
[13] Verb infinite mood	**.718***	**.488***	**.644***	**.481***	**.649***	**.440***	**.752***	**.528***	**.720***	**.459***	**.554***	**.753***	**.395***	**.405**
[14] Verb gerundive mood	**.574***	__	**.585***	__	**.554***	__	**.691***	-.038	**.677***	__	**.499***	__	**.519***	**.558***
[15] Verb participle mood	**.530***	.210	**.439***	**.395***	**.380***	**.323**	**.554***	**.335***	**.349***	**.368***	**.527***	.204	**.371***	.148
[16] Verb indicative mood	**.584***	.223	**.630***	**.422***	**.581***	.100	**.697***	**.344***	**.675***	**.323**	**.686***	**.495**	**.491***	.156
[17] Verb present tense	**.573***	**.254**	**.622***	**.307**	**.574***	**.275**	**.683***	**.394***	**.558***	**.296**	**.599***	**.568**	**.727***	**.527**
[18] Verb imperfect tense	**.741***	**.638***	**.786***	**.533***	**.768***	**.635***	**.849***	**.542***	**.771***	**.479***	**.813***	**.884***	**.777***	**.432**
[19] Verb past tense	**.703***	.214	**.832***	.088	**.787***	.080	**.840***	**.260***	**.811***	.187	**.902***	__	**.801***	**.504**
[20] Main clause	**.492***	.215	**.395***	**.198**	**.495***	.046	**.520***	**.215**	**.518***	.191	**.337**	__	**.277**	.097
[21] Subord. Clause	**.492***	.215	**.478***	**.204**	**.495***	.151	**.520***	**.209**	**.518***	**.254**	**.337**	.145	**.277**	.238
[22] Embedded subord. clause	**.356***	**.303**	**.547***	**.351***	**.369***	**.323**	**.529***	**.415***	**.463***	**.404***	**.422**	**.472**	**.499***	.173
[23]Prepositional 'chains'	**.647***	**.352**	**.567***	.305	**.679***	.225	**.740***	**.424***	**.627***	**.514***	**.724***	**.712***	**.664***	**.507**
[24] Length of depend. links	**.608***	**.403***	**.582***	**.431***	**.457***	**.278**	**.619***	**.433***	**.571***	**.468***	**.498***	.215	**.512***	**.562***
[25] Longest depend.links	**.643***	**.321**	**.586***	**.345***	**.523***	**.307**	**.621***	**.428***	**.599***	**.493***	**.514***	.160	**.578***	**.596***
[26] Parse tree depth	**.559***	.166	**.518***	**.275**	**.506***	**.280**	**.671***	**.379***	**.602***	**.405***	**.509***	.376	**.499***	.294
[27] Verb arity	**.630***	.231	**.583***	**.236**	**.417***	.191	**.588***	**.365***	**.548***	**.321**	**.494**	.019	**.511***	.003
[28] Verbal roots with subject	**.469***	.182	**.570***	**.324***	**.438***	**.331**	**.585***	**.347***	**.473***	**.365***	.017	.439	**.614***	.216
[29] Post-verbal objects	**.566***	.224	**.524***	.178	**.471***	**.288**	**.634***	**.389**	**.575***	**.228**	**.573***	.162	**.511***	.082

Features	Insert		Delete		Reordering		LexSub word		LexSub phrase		Split		Insert Verb	
[30] Pre-verbal objects	**.416***	**.340**	**.381***	**.227**	**.380***	**.605***	**.616***	**.307***	**.519***	**.315**	**.670***	-.076	**.619***	-.065
[31] Post-verbal subj	**.363***	.204	**.498***	**.294**	**.207**	**.500***	**.521***	**.349***	**.266***	**.228**	**.615***	**.570**	**.344***	.343
[32] Pre-verbal subj	**.476***	.141	**.534***	.163	**.220**	.076	**.568***	**.326***	**.328***	**.324**	**.441**	.089	**.572***	-.024
[33] Post-verbal subord. Clause	**.552***	**.337**	**.387***	**.336***	**.488***	**.260**	**.647***	**.469***	**.528***	**.388***	**.505***	.556	**.385***	.052
[34] Pre-verbal subord. clause	**.299***	.155	**.592***	**.233**	**.445***	.105	**.495***	**.159**	**.308***	.085	**.315**	.444	**.424***	-.100
[35] Clause length	**.707***	**.485***		**.481***	**.635***	**.388**	**.711***	**.513***	**.650***	**.450***	**.637***	.514	**.622***	**.462**

References

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. System, 37(4): 585–599.

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero E. G. and de Mattos Fortes, R.P. (2008). Towards Brazilian Portoguese automatic text simplification systems. Proceeding of the eighth ACM symposium on Document engineering, 240–248.

Anderson, W. (2006). *The phraseology of administrative French: a corpus-based study.* Series: Language and computers: studies in practical linguistics, 57. Rodopi.

Aranzabe, M. J., De Ilarraza, A. D., Gonzalez-Dios, I. (2013). Transforming complex sentences using dependency trees for automatic text simplification in Basque. Procesamiento del lenguaje natural, 50, 61–68.

Arosio, F., Adani, F., & Guasti, M. T. (2005). Processing grammatical features by Italian children. In A. Belletti, E. Bennati, C. Chesi & I. Ferrari (Eds.), Acquisition and development. Proceedings of GALA 2005 (pp. 15-27). Siena: Cambridge Scholars Press.

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), New York City, New York, pages 166–170.

Baker, M. (1988). Incorporation: A Theory of Grammatical Function Changing, Chicago University Press.

Barlacchi, G. and Tonelli, S.(2013). ERNESTA: A Sentence Simplification Tool for Children’s Stories in Italian. Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2013), 476–487.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. In Computational Linguistics, 34(1), pp. 1–34.

Bastiaanse, R. , Rispens, J., Ruigendijk, E., Juncos Rabadan, O. and Thompson, C.K. (2002). Verbs: Some properties and their consequences for agrammatic Broca’s aphasia. Journal of Neurolinguistics, 15: 239-264.

Brouwers, L., Bernhard, D., Ligozat, A.-L. and François, T (2014). Syntactic Sentence Simplification for French. Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden:47–56.

Belletti, A. and Rizzi, L. (1988). Psych-Verbs and Th-Theory, *Natural Language and Linguistic Theory*, 6, 3, 1988, pp.291-352.

Belletti, A. (Eds.) (2002). Structures and Beyond. The Cartography of Syntactic Structures, Vol. 3, Oxford: Oxford University Press.

Belletti, A. and Contemori, C. (2010). Intervention and attraction on the production of subject and object relatives by Italian (young) children and adults. In J. Costa, A. castro, M. Lobo, and F. Pratas (eds), Proceedings of Gala, 2009, Cambridge: Cambridge Scholars Publishing.

Belletti A., Chesi, C. (2011) Relative clauses from the input: syntactic considerations on a corpus-based analysis of Italian. STiL - Studies in Linguistics. 4:5-24.

Belletti, A., Rizzi, L. (2012). Moving Verbal Chunks in the low functional field, In: Brugè L., A. Cardinaletti, G.Giusti, N.Munaro, C.Poletto, (eds.), Functional Heads, Oxford University Press, 129-137.

Belletti, A. & Rizzi, L.(2013). Intervention in grammar and processing. In Ivano Caponigro & Carlo Cecchetto (eds.) From Grammar to Meaning: The Spontaneous Logicality of Language. Cambridge: Cambridge University Press, 293 - 311.

Berruto, G. (1987). Sociolinguistica dell'italiano contemporaneo. Roma: La Nuova Italia Scientifica.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In Hayes, J. R. (Eds.), *Cognition and the Development of Language*, Wiley, New York;

Bianchi, V. (2002). Headed relative Clauses in Generative Syntax. Part I. Glot International 6(7):197–204

Biber, D. (1993). Using Register–diversified Corpora for General Language Studies. Computational Linguistics Journal, 19(2): 219–241.

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

Biber, D. (1988). Variation across speech and writing. Cambridge & New York, Cambridge University Press.

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison, Cambridge & New York, Cambridge University Press.

Blache, P. (2011). A computational model for linguistic complexity, in Proceedings of the first International Conference on Linguistics, Biology and Computer Science.

Boland, J. E., Tanenhaus,M. K., & Garnsey, S. (1990). Evidence for the immediate use of verb control information in sentence processing. Journal of Memory and Language, 29, 413–432.

Bosco, C., Dell'Orletta, F., Montemagni, S., Sanguinetti, M. and Simi, M. (2014). The Evalita 2014 Dependency Parsing Task. Proceedings of Evalita'14, Evaluation of NLP and Speech Tools for Italian, Pisa, December.

Bott, M. and Saggion, H. (2014). Text simplification resources for Spanish. Language Resources and Evaluation, 48(1): 93–120.

Bonin, F., Dell'Orletta, F., Montemagni, S., Venturi, G., (2010). A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora, in Proceedings di LREC'10 - Seventh International Conference on Language Resources and Evaluation, Valletta (Malta), 17-23 May 2010, pp. 3222 - 3229.

Brennan, J. and Pylkkänen, L. (2010). Processing Psych Verbs: Behavioral and MEG Measures of Two Different Types of Semantic Complexity. Language and Cognitive Processes, 25 (6), 77-807.

Brunato D., Dell'Orletta F., Venturi G., Montemagni S. (2015). Design and Annotation of the First Italian Corpus for Text Simplification. In Proceedings of the 9th Linguistic Annotation Workshopthe (LAW'15), 5 June, Denver, Colorado, USA.

Burani, C. & Caramazza, A. (1987). Representation and processing of derived words. Language and Cognitive Processes, 3, 217-227.

Burani, C., Barca, L., Arduino, L.S. (2001). Una base di dati sui valori di età di acquisizione, frequenza, familiarità, immaginabilità, concretezza, e altre variabili lessicali e sublessicali per 626 nomi dell'italiano, in Giornale italiano di psicologia 4/2001, pp. 839-856.

Calvino, I. Per ora sommersi dall'antilingua, Il Giorno, 8 february, 1965.

Caramazza, A. and Zurif, E. (1976). Dissociations of algorithmic and heuristic processes in sentence comprehension: Evidence from aphasia. Brain and Language, 3, 572-582.

Carroll, J., Minnen, G. , Canning, Y., Devlin, S. and Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, Association for the Advancement of Artificial Intelligence (AAAI).

Caseli, H., Pereira, T., Specia, L., Pardo, T., Gasperin, C., and Aluísio, S. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. Proceedings of the 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), March 0107, Mexico City.

Chall, J., Dale, E. (1995). Readability revisited: the new Dale-Chall readability formula, Cambridge, Mass: Brookline Books.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. Proceedings of the international conference on computational Linguistics, 1041–1044.

Chesi, C. (2012). Competence and Computation. Toward a Processing Friendly Minimalist Grammar. UNIPRESS, Padova.

Chesi, C. and Moro, A. (2014). Computational complexity in the brain, in Frederick J. Newmeyer and Laurel B. Preston (eds.), Measuring Linguistic Complexity. Oxford: Oxford University Press.

Chomsky, C. (1969). The Acquisition of Syntax in Children from 5 to 10, Research Monograph No. 57, London, The MIT Press.

Chomsky, N. (1957). *Syntactic Structures*, The Hague/Paris: Mouton.

Chomsky, N. and Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, eds., Handbook of mathematical psychology, vol. 2, 269-321. New York: Wiley.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, Cambridge, Massachusetts: MIT Press

Chomsky, N. (1981). Lectures on Government and Binding: The Pisa Lectures. Holland: Foris

Chomsky, N. (1995). The Minimalist Program. Cambridge, MA: MIT Press.

Chomsky, N. (2001). Derivation by phase. In: Michael Kenstowicz (Eds.). Ken Hale: a life in language. Cambridge, MA: MIT Press, 1-53.

Ciccarelli, L. and De Vincenzi, M. (1996), Lo studio della complessità linguistica in rapporto alle strategie cognitive di analisi del linguaggio, In A. Colombo e W. Romani (edited by), *E' la lingua che ci fa uguali*, Firenze, La Nuova Italia.

Ciccarelli, L. (1998). *Comprensione del linguaggio, dei processi di elaborazione e memoria di lavoro: uno studio in et`a prescolare*. PhD dissertation, University of Padua.

Cinque, G. (Eds.), (2002). *The Structure of CP and DP: The Cartography of Syntactic Structures*, vol. 1. Oxford University Press, Oxford, New York.

Cipriani, P., Chilosi, A. M. , Bottari, P. and Pfanner, L. (1993*). L'acquisizione della morfosintassi in italiano: fasi e processi*. Padova: Unipress.

Collins-Thompson, K. (2014), Computational assessment of text readability: A survey of current and future research, *International Journal of Applied Linguistics*.

Coltheart, M. (1981). *The MRC Psycholinguistic Database*. Quarterly Journal of Experimental.

Coltheart, M., Davelaar, E., Jonasson, J. F., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Eds.), Attention & Performance VI (pp. 535-555). Hillsdale, NJ:Erlbaum. Domínguez, A.,

Conrad, Susan and Biber, Douglas (eds.) (2001).Variation in English: Multi-dimensional studies. Harlow, England: Longman.

Cortelazzo, M. A. (1990). *Lingue speciali. La dimensione verticale*. Padova, Unipress.

Cortelazzo, M. A. – Pellegrino, Federica (2003). Guida alla scrittura istituzionale, Roma-Bari, Laterza.

Cortelazzo, M. A. (1999). *Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini*, Padova, Comune di Padova.

Cortelazzo, M.A. (2005). *Il Comune scrive chiaro. Come semplificare la comunicazione al cittadino. Con 24 esempi di testi rielaborati e le istruzioni per scrivere con stile*, Santarcangelo di Romagna, Maggioli.

Cutler, A. (1983). *Lexical Complexity and Sentence Processing*, in Flores d'Arcais, G. & R. Jarvella, The Process of Language Understanding, Wiley, N.Y.

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19, 450–466.

Dardano, M. (1973). Il linguaggio dei giornali italiani, Bari, Laterza.

Davison, A. & Green, G. M. (Eds), (1988). Linguistic complexity and text comprehension: Readability issue reconsidered. Hillsdale, N J: Erlbaum.

Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. Reading Research Quarterly, 17, 187-209.

De Belder, J. and Moens, M.F. (2010). Text Simplification for Children. Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems.

De Belder, J., Deschacht, K. and Moens, M.F. (2010). Lexical simplification. Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication.

De Mauro, T. (2000). Il dizionario della lingua italiana. Paravia, Torino.

De Vincenzi, M. (1991). Syntactic Parsing Strategies in Italian, Dordrecht, Kluwer.

De Vincenzi, M. and Di Matteo, R. (2004). *Come il cervello comprende il linguaggio*. Laterza, Bari.

Dell'Orletta F., Montemagni S., Venturi G. (2011). *READ-IT: assessing readability of Italian texts with a view to text simplification*. In: SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (Edimburgo, UK, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

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, December.

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

Dell'Orletta, F., Montemagni, S., 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), 7-11 September, Hissar, Bulgaria, pp. 189-197.

Dipartimento della Funzione Pubblica (1993). Codice di stile delle comunicazioni scritte ad uso delle pubbliche amministrazioni. Roma: Istituto Poligrafico e Zecca dello Stato.

Dipartimento della Funzione Pubblica per l'efficienza delle amministrazioni (2002), Direttiva sulla semplificazione del linguaggio dei testi amministrativi.

Drndarevi´c, B., Stajner, S., Bott, S., Bautista S., and Saggion, H. (2013). Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. A. Gelbukh (ed.) 14th Conference on Computational Linguistics and Natural Language Processing (CICLing'14), LNCS 7817 (2):488–500.

Drndarevi´c, B., Stajner, S., and Saggion, H. (2012). Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. Proceedings of "Easy to read on the web", online symposium.

DuBay, W. H. (2006). The classic readability studies, Impact Information Costa Mesa.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: The MIT press.

Feng, L., Elhadad N. and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), pp. 229–237.

Ferguson, C.A. & De Bose C. E. (1977). Simplified registers, broken language and pidginization. In Valdman, A. (Eds.) Pidgin and Creole Linguistics. Bloomington: Indiana University Press. Pp.99-125.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. Cognitive Psychology, 47, 164-203.

Fiorin, G. (2009). The Interpretation of Imperfective Aspect in Developmental Dyslexia. Proceedings of the 2nd International Clinical Linguistics Conference, Universidad Autonoma de Madrid, Universidad Nacional de Educacìon a Distancia, and Euphonia Ediciones.

Fioritto, A., (1997). Manuale di stile, Bologna, il Mulino.

Flesh R. (1948). A New Readability Yardstick, Journal of Applied Psychology, Vol. 32(3), Jun 1948

Fodor, J. (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J., Bever, T. and Garrett, M. (1974) *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar*. New York: McGraw Hill.

Fodor, J. D., Garrett, M. F., & Bever, T. G. (1968). Some syntactic determinants of sentential complexity. Perception and Psychophysics II: Verb structure(3), 453-461.

Fortis, D . (2005). Il linguaggio amministrativo italiano, in Revista de Liengua i dret, n.43, pp. 47-116.

Franceschini, F. and Gigli, S. (edited by), (2003). Manuale di scrittura amministrativa, Roma: Agenzia delle Entrate.

François, T. and Fairon, C. (2012). An "AI readability" formula for French as a foreign language In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012), Jeju, 466-477.

Fraser, C. K, Meltzer, J. A., Graham, N.L, Leonard, C., Hirst, C., Black, S.E. and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex, 55:43–60.

Frazier, L. and Fodor, J.D. (1978). The sausage machine: A new two-stage parsing model. Cognition 6, 291-325.

Frazier, L. (1979). On Comprehending Sentences: Syntactic parsing strategies. Ph.D. Dissertation, University of Connecticut.

Frazier, L., Clifton, C., Jr., & Randall, J. (1983). Filling gaps: Decision principles and structure in sentence comprehension. Cognition, 13, 187–222.

Frazier, L. (1985). Syntactic Complexity, in D. Dowty, L. Karttunen and A. Zwicky, (eds.) Natural Language Parsing: Psychological, Computational and Theoretical Perspective, Cambridge University Press, 129-89.

Frazier, L. (1987). Sentence processing: a tutorial review. In Attention and performance XII: The Psychology of Reading, 559–586. Erlbaum.

Friedmann N., Belletti, A., Rizzi, L. (2009). Relativized Relatives: Types of intervention in the acquisition of A'-dependencies. LINGUA, vol. 119, pp. 67-88.

Friedmann, N. and Shapiro, L. P. (2003). Agrammatic comprehension of simple active sentences with moved constituents: Hebrew OSV and OVS structures. Journal of Speech Language and Hearing Research, 46:441–463.

Friedmann, N. and Y. Grodzinsky, Y. (1997). Tense and agreement in agrammatic production: pruning the syntactic tree, Brain and Language, 56: 397-425.

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

Gibson, E., Pearlmutter, N., Canseco Gonzalez, E., Hickock,G. (*1996*). Recency preference in the human sentence processing mechanism. Cognition, 59, 23–59.

Gildea, D. (2001). Corpus variation and parser performance. Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001), Pittsburgh, PA.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. Journal of Experimental Psychology: Learning, Memory and Cognition, 27, 1411–1423.

Gordon, P.C., & Hendrick, R. (1998). The representation and processing of coreference in discourse. Cognitive Science, 22(4), 389-424.

Gordon, P.C., Grosz, B.J., and Gilliom, L.A. (1993). Pronouns, names and the centering of attention in discourse. Cognitive Science, 17, 311-347.

Gordon, P. C., Randall H. and Johnson M. (2004). Effects of noun phrase type on sentence complexity. Journal of Memory and Language. 51: 97-114.

Gotti, M. (2005), Investigating Specialized Discourse. Bern: Peter Lang.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers, 36, 193-202.

Grillo, N. (2008). Generalized Minimality – Syntactic Underspecification in Broca's Aphasia. PhD Dissertation, University of Utrecht, University of Siena. Distributed by LOT.

Grodzinsky, Y. (1990). Theoretical perspectives on language deficit. Cambridge, MA: MIT Press

Grosz, B. J., Aravind, K. J. and Weinstein, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. In Computational Linguistics 2(21), pp. 203-225.

Grosz, B., J., Aravind, K. J. and Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In Proceedings of 21st Annual Meeting of the ACL, pages 44-50. Association of Computational Linguistics.

Guasti, M.T. (2002) Language acquisition. The growth of grammar, MIT press.

Haliday, M.A.K., Hasan, R., (1976). Cohesion in English. Longman, London.

Halliday, M. A. K. (1985). *Spoken and Written Language*. *Geelong, Vic.: Deakin University Press*.

Halteren, V.H. (2004). Linguistic profiling for author recognition and verification. Proceedings ACL 2004.

Harrington, M. (2001). Sentence processing, Cognition and second language instruction, 91-124.

Hasbrouck, J. E., Tindal, G., & Parker, R. I. (1994). Objective procedures for scoring students' writing. Teaching Exceptional Children, 26(2),18-22.

Inui, K., Fujita, A., Takahashi, T., Iida, R. and Iwakura, T.(2003). Text Simplification for Reading Assistance: A Project Note. Proceedings of the Second International Workshop on Paraphrasing, ACL.

Istituto di Teorie e Tecniche dell'informazione giuridica (ITTIG) e Accademia della Crusca (2011). Guida alla redazione degli atti amministrativi. Regole e suggerimenti.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. Cognitive Psychology, 13, 278-305.

Just, M. A. & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension, Psychological Review 87(4), 329-354.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F. and Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. Science, 274(5284), 114-116.

Kate, R. J., Luo, X., Patwardhan S., Franz M., Florian R., Mooney R.J., Roukos S., and Welty, C.(2010). Learning to predict readability using diverse linguistic features. In proceedings of the 23rd International Conference on Computational Linguistics, pages 546–554.

Kennison, S.M., and Gordon, P.C. (1997). Comprehending referential expressions during reading: Evidence from eye tracking. Discourse Processes, 24, 229-252.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. Cognition, 2(1), 15–47.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. Journal of memory and language, 30(5): 580-602.

Kintsch, W. and Keenan, J. (1973). Reading rate and retention as a function of the number of prepositions in the base structure of sentences. Cognitive Psychology, 5: 257–274.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. Psychological Review, 95, 163-182.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25:259–284.

Lapata, M. and Barzilay, R. (2005). Automatic Evaluation of Text Coherence: Models and Representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1085-1090. Edinburgh.

Laudanna, A. and Voghera, M. (Eds.), (2006). Il linguaggio. Strutture linguistiche e processi cognitivi. Bari: Laterza.

Levin, B. (2006). Determining semantic prominence and argument realization: themes, recipients, spatial goals, and dative verbs. Handout VII from the course on "Lexical Semantics and Argument Realization", DGfS/GLOW Summer School, Stuttgart.

Lin, S.Y., Su, C.C., Lai, Y.D., Yang, L.C., & Hsieh, S.K. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet. International Journal of Computational Linguistics and Chinese Language Processing.

Lubello, S. (2014). Il linguaggio burocratico, Roma, Carocci.

Lucisano, P., Piemontese, M. E. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. In: «Scuola e città», 34.

Manouilidou, C., de Almeida, R.G., Schwartz, G., & NPV Nair (2009). Thematic Roles in Alzheimer's Disease: Hierarchy Violations in Psychological Predicates. Journal of Neurolinguistics 22(2): 167-186.

Maratsos, M. (1974). Children who get worse at understanding the passive: A replication to Bever. Journal of Psycholinguistic Research, 3:65–74.

Marinelli, C.V., Traficante, D., Zoccolotti, P., Burani, C. (2013). Orthographic Neighborhood-Size Effects on the Reading Aloud of Italian Children With and Without Dyslexia, *Scientific Studies of Reading* Vol. 17, Iss. 5.

McClelland, J. L. (1988). Connectionist models and psychological evidence. Journal of Memory and Language, 27, 107-123.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101, 676-703

McDonald, R. and Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNNL), 122-131.

McWhorter, J. (2001). The world's simplest grammars are creole grammars. Linguistic Typology 6: 125-166.

Merlini Barbaresi, L. (Eds.) (2003). *Complexity in Language and Text*. Pisa: Plus - Pisa University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63 (2): 81–97.

Miller, G. A. (1990). WordNet: An on-line lexical database. International Journal of Lexicography, 3(4):235–312. Special Issue.

Miller, G. A., and Mckean, K. O. (1964). A chronometric study of some relations between sentences. Quart. J. exp. Psychol., 16, 297-308.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F. and Delmonte, R. (2003). *Building and the Italian Syntactic-Semantic Treebank*, pages 189–210. Kluwer, Dordrecht.

Montemagni, S. (2013), Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. In *Studi Italiani di Linguistica Teorica e Applicata* (SILTA), Anno XLII, Numero 1, pp. 145-172.

Narayan, S. and Gardent, D. (2014). Hybrid Simplification using Deep Semantics and Machine Translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 435–445.

Newmeyer ,F. and Preston, L. (eds). (2014) Measuring Linguistic Complexity. Oxford: Oxford, University Press.

Ozuru, Y., Dempsey, K. and McNamara, D. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. Learning and Instruction, (19): 228–242.

Partee, B. and Rooth, M. (1983). Generalized Conjunction and Type Ambiguity. In R. Bäuerle, C. Schwarze, and A. von Stechow (eds.), Meaning, Use and Interpretation of Language, de Gruyter, Berlin.

Perea, M., & Rosa, E.. (2000). Repetition and form priming interact with neighborhood density at a brief stimulus onset asynchrony. Psychonomic Bulletin and Review.

Perfetti, C. A. (1985). Reading Ability. Oxford: Oxford University Press.

Pesetsky, D. (1995) Zero Syntax, MIT Press, Cambridge Mass.

Petersen, S. E. and Ostendorf, M. (2007). Text Simplification for Language Learners: A Corpus Analysis. Speech and Language Technology for Education (SLaTE).

Pianta, E., Bentivogli, L. and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In First International Conference on Global WordNet, pp. 292–302, Mysore, India.

Pickering, M.J. and *van Gompel*, R.P.G.. (2006). Syntactic parsing. In M.J. Traxler. & M. A. Gernsbacher (Eds.), Handbook of. Psycholinguistics, 2nd Ed. (pp.455-503). Amsterdam: Academic Press.

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

Pikulski, J.J., 2002. Readability. U.S.A: Houghton Mifflin Company.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of EMNLP-CoNLL.

Pustejovsky, J. (1995). The Generative Lexicon, MIT Press, Cambridge, MA.

Quochi, V. (2007). A Usage-Based Approach to Light Verb Constructions in Italian: Development and Use, PhD dissertation, University of Pisa.

Raso, T. (2005). La scrittura burocratica. La lingua e l'organizzazione del testo, Roma, Carocci.

Rello, L. and Baeza-Yates, R. (2013). Good Fonts for Dyslexia. In ASSETS 2013: The 15th International ACM SIGACCESS Conference of Computers and Accessibility, Bellevue, Washington USA, 22-24 October.

Richards, B. J. (1987). Type/token ratios: what do they really tell us?, *Journal of Child Language 14*: 201-209.

Rizzi, L. (1990). Relativized Minimality, MIT Press, Cambridge, Mass.

Rizzi, L. (1997).The Fine Structure of the Left Periphery, in L.Haegeman (Eds.) Elements of Grammar. Amsterdam: Kluwer Grammar. Dordrecht: Kluwer, 281-337.

Rizzi, L. (1998), On the Study of Language as a Cognitive Capacity, Plenary Lecture, in B. Caron, ed., Proceedings of the XVIth International Congress of Linguists (CDRom), Pergamon, Elsevier Sciences, New York.

Rizzi, L. (2001). Relativized Minimality Effects. In: Mark Baltin & Chris Collins (eds). The Handbook of Contemporary Syntactic Theory. Oxford: Blackwell, 89-110.

Rizzi, L. (Eds.) (2002). The Structure of CP and IP, The Cartography of Syntactic Structures, vol. 2. Oxford: Oxford University Press.

Rizzi, L. (2003), Some Elements of the Study of Language as a Cognitive Capacity, in Dimitri, ed., N., M. Basili, I. Gilboa, (eds.), Cognitive Processes and Economic Behaviour, Routledge, London and New York, 104-136.

Rizzi, Luigi. (Eds.) (2004). The Structure of CP and IP. The Cartography of Syntactic Structures, vol.2, New York: Oxford University Press

Rizzi, L. (2013). Locality, *Lingua* 130, pp. 169-186.

Roark, B., Mitchell, M., Hollingshead K. (2007). Syntactic complexity measures for detecting mild cognitive impairment, in Proc. ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP'07), pp.1-8.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structures of categories. Cognitive Psychology, 7, 573-605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8, 382-439.

Rosenbaum, P.S. (1967). The grammar of English predicate constructions. MIT Press, Cambridge Mass.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Eds.), Attention and performance, 6 (pp. 573-603). Hillsdale, NJ: Erlbaum.

Sabatini, F. (1990) Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi, in AA.VV, Corso superiore di studi legislativi 1988- 1989. Padova, CEDAM-Casa Editrice dott. Antonio Milani, 675-724.

Sahakian, S. and Snyder, B. (2012). Automatically learning measures of child language development. Proceedings of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 95–99.

Sagae, K., Lavie, A., Macwhinney, B.(2005). Automatic measurement of syntactic development in child language, in Proceedings of the 43rd Annual Meeting of the ACL.

Sanders, T. and Noordman, L. (2000) The Role of Coherence Relations and their Linguistic Markers in Text Processing, Discourse Processes 29(1): 37–60.

Scarton, C., Almeida, D., M., Aluísio, S. (2009). Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natu-ral: adaptando as métricas do Coh-Metrix para o Português. In Proceedings of STIL-2009, São Carlos, Brazil.

Scott, Cheryl M. (2009). A case for the sentence in reading comprehension. Language, Speech, and Hearing Services in Schools, 40.2: 184-191.

Searle, John R. (1972). Chomsky's revolution in linguistics. The New York Review of Books, June 29, 1972.

Segui J., Mehler J., Frauenfelder U., Morton J. (1982). The word frequency effect and lexical access. Neuropsychologia 20: 615–627.

Seidenherg, M. S. (1989). Reading complex words. In G. N. Carlson & M. K. Tanenhaus (Eds.), Linguistic structure in language processing (pp. 53-105). Amsterdam: Reidel.

Serianni, L. (2005), Un treno di sintomi. I medici e le parole. Percorsi linguistici nel passato e nel presente, Milano, Garzanti.

Shapiro, L.P., Zurif, E., and Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. Cognition, 27, 219-246.

Sheehan, K.M., Flor, M., and Napolitano, D. (2013). A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity. In Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA), Atlanta, GA.

Siddharthan, A. (2014). A survey of research on text simplification. International Journal of Applied Linguistics, 165(2): 259–298.

Siddharthan, A. and Angrosh, M.A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014).

Siddharthan, (2011). Text Simplification Using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. Proceedings of the 13th EuropeanWorkshop on Natural Language Generation (ENLG'11), Nancy, France: 2–11.

Siddharthan, A. (2010). Complex lexico-syntactic reformulation of sentences using typed dependency representations. Proceedings of the 6th International Natural Language Generation Conference (INLG). Association for Computational Linguistics, 125-133.

Siddharthan, A. (2002). An Architecture for a Text Simplification System. Proceedings of the Language Engineering Conference (LEC 2002).

Silva, M., Abchi, V. & Borbone, A. (2010). Subordinate Clauses Usage and Assessment of Syntactic Maturity: a comparison of oral and written retellings in beginning writers. Journal of Writing Research 02.1. 47-64.

Slobin, D. I. (1966). Grammatical transformations and sentence comprehension in childhood and adulthood. Journal of Verbal Learning and Verbal Behavior, 5, 219–227.

Slobin, D. I. and Bever, R. G. (1982). Children use canonical sentence schemas. A cross-linguistic study of word order and inflections. Cognition, 12(3): 229–265.

Sobrero, A.A. (1993). «Lingue speciali». In Id. (edited by). Introduzione all'italiano contemporaneo. La variazione e gli usi. Roma-Bari: Laterza, pp. 237-277.

Sorace, A. (1993). Incomplete vs. divergent representations of unaccusativity in non native grammars of Italian. Second Language Research, 9(1), 22–47.

Specia, L. (2001). Translating from complex to simplified sentences. Computational Processing of the Portoguese Language, 6001:30–39.

Stavrakaki, S. (2001). Comprehension of reversible relative clauses in specifically language impaired and normally developing Greek children. Brain and Language, 77, 419-431.

Stowe, L. (1989). Thematic structures and sentence comprehension. In G. N. Carlson and M. K. Tanenhaus (Eds.), Linguistic structure in language processing (pp. 319-356). Dordrecht, The Netherlands: Kluwer.

Szmrecsányi, Benedikt. (2004). On operationalizing syntactic complexity. In G.Purnelle; C. Fairon; and A. Dister (Eds.). Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la- Neuve: Presses universitairs de Louvain. 2: 1032-1039.

Tanenhaus, M.K. & Trueswell, J.C. (1995). Sentence Comprehension. In Eimas & Miller (Eds.) Handbook in Perception and Cognition, Volume 11: Speech Language and Communication. Academic Press, pp. 217-262.

Tessuto, G. (2006). Simplifying Italian Administrative Language: An Overview, In Wagner A., Cacciaguidi-Fahy S. (Dds) (2006) Legal Language and the Search for Clarity: Practice and Tools. London: Peter Lang – Linguistic Insights.

Thompson, C.K., Shapiro, L.P. (2007). Complexity in treatment of syntactic deficits. *American Journal of Speech-Language Pathology*, 16:30–42.

Thorndike, E. L. (1921). The teacher's word book. New York: Teacher's College, Columbia University.

Tonelli, S., Manh, K.T. and Pianta, E. (2012). Making Readability Indices Readable, in Proceedings of the NAACL Workshop on Predicting and Improving Text Readability for Target Reader Population (PITR), Montreal, Canada.

Traxler, Matthew J., Robin K. Morris & Rachel E. Seely (2002). Processing subject and object relative clauses: Evidence from eye movements. Journal of Memory & Language, 47(1), 69–90.

Ungari, P. (1983), in Atti del Convegno, Il Linguaggio della divulgazione, II convegno, Milano, Selezione del Reader's Digest.

Vacca, R, Franchina, V. (1986). Taratura di indici di Flesch su testo bilingue italiano inglese di unico autore. «Linguaggi», 3: 47-49

Van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Venturi, G. (2013). Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach. *The International Journal of Speech, Language and the Law (IJSLL)*, (To appear).

Venturi, G. (2012). Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts. In: Proceedings of the LREC 2012 4th Workshop on "Semantic Processing of Legal Texts", pp. 1-12, Istanbul, Turkey, 27 May.

Viale, M. (2008) Studi e ricerche sul linguaggio amministrativo, Padova, Cleup.

Voghera, M. (2001), Riflessioni su semplificazione, complessità e modalità di trasmissione: sintassi e semantica, in Maurizio Dardano et al. (edited by), Scritto e parlato. Metodi, testi e contesti, Atti del Colloquio Internazionale di Studi, Aracne, Roma, pp.65-78.

Voghera, M. (2005). La misura delle categorie sintattiche, in Isabella Chiari – Tullio De Mauro (edited by), Parole e numeri. Analisi quantitative dei fatti di lingua, Aracne, Roma, pp.125-138.

Wanner, Eric & Michael Maratsos (1978). An ATN approach to comprehension. In Halle, M., Bresnan, J. & Miller, G. A. (eds.) Linguistic Theory and Psychological Reality. pp. 119–161. Cambridge, MA: MIT Press.

Warren, T., Gibson, E. (2002). The influence of referential processing on sentence complexity. Cognition 85, 79–112.

Wilcock, G. (2009). Introduction to Linguistic Annotation and Text Analytics. *Synthesis Lectures on Human Language Technologies,* Morgan & Claypool.

Wray, D. & Janan, D. (2013). Exploring the Readability of Assessment Tasks: The Influence of Text and Reader Factors. Multidisciplinary Journal of Educational Research, 3(1), 69-95.

Victor, H. (1960). A model and an hypothesis for language structure. Proceedings of the American Philosophical Society, 104:444-466.

Volpato, F. (2010). The acquisition of relative clauses and phi-features: evidence from hearing and hearing-impaired populations. PhD dissertation, Ca' Foscari University of Venice.

Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 409–420.

Zipf, G. K. (1935). *The Psychobiology of Language*. Houghton Mifflin, Boston.

Zhu, Z., Bernhard, D. and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. Proceedings of the 23rd international conference on computational linguistics, Association for Computational Linguistics. 1353–1361.