Lexicon and Syntax: Complexity across Genres and Language Varieties

Pietro dell'Oglio•, Dominique Brunato◇, Felice Dell'Orletta◇
• University of Pisa
pietrodelloglio@live.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

English. This paper presents first results of an ongoing work to investigate the interplay between lexical complexity and syntactic complexity with respect to nominal lexicon and how it is affected by textual genre and level of linguistic complexity within genre. A cross-genre analysis is carried out for the Italian language using multi–leveled linguistic features automatically extracted from dependency parsed corpora.

Italiano. *Questo articolo presenta i primi risultati di un lavoro in corso volto a indagare la relazione tra complessità lessicale e complessità sintattica rispetto al lessico nominale e in che modo sia influenzata dal genere testuale e dal livello di complessità linguistica interno al genere. Un'analisi comparativa su più generi è condotta per la lingua italiana usando caratteristiche linguistiche multilivello estratte automaticamente da corpora annotati fino alla sintassi a dipendenze.*

1 Introduction

Linguistic complexity is a multifaceted notion which has been addressed from different perspectives. One established dichotomy distinguishes a "global" vs a "local" perspective, where the former considers the complexity of the language as a whole and the latter focuses on complexity within each sub-domains, i.e. phonology, morphology, syntax, discourse (Miestamo, 2008). While measuring global complexity is a very ambitious and probably hopeless endeavor, measuring local complexities is perceived as a more doable task (Kortmann and Szmrecsanyi, 2012). The level of complexity within each subdomains indeed has been formalized in terms of distinct parameters that capture either internal properties of the language (in the "absolute" notion of complexity) or phenomena correlating to processing difficulties from the language user's viewpoint (in the "relative" notion of complexity) (Miestamo, 2008). For instance, complexity at lexical level has been computed in terms of *length* (measured in characters or syllables), of *frequency* either of the whole surface word (Randall and Wayne, 1988; Chiari and De Mauro, 2014) or of its internal components (see e.g. the *root frequency effect* (Burani, 2006)), *ambiguity* and *familiarity*, among others. At syntactic level, much attention has been paid on canonicity effects due to word order variation (Diessel, 2005; Hawkins, 1994; Futrell et al., 2015), as well as on long-distance dependencies (Gibson, 1998; Gibson, 2000) proving their effect on a wide range of psycholinguistic phenomena, such as the subject/object relative clauses asymmetry or the garden path effect in main verb/reduced–relative ambiguities.

An interesting question addressed by recent corpus-driven research is how language complexity is affected by textual genre. At syntactic level, the study by Liu (2017) on ten genres taken from the British National Corpus showed that genre-specific stylistic factors have an influence on the distribution of dependency distances and dependency direction. Similarly for Italian, Brunato and Dell'Orletta (2017) investigated the influence of genre, and level of complexity within genre, on a range of factors of syntactic complexity automatically computed from dependency-parsed corpora. Inspired by that work, we also intend to analyze the effect of genre on linguistic complexity. However, unlike the dominant local approach, where each subdomain is typically studied in isolation, our contribution intends to address the interrelation between different levels, i.e. lexicon and syntax. Specifically, we investigate the fol-

lowing questions:

- to what extent is lexical complexity influenced by genre?
- to what extent is lexical complexity influenced by the level of complexity within the same genre?
- is there a correlation between lexical complexity and syntactic complexity? Does it vary according to genre and level of complexity within the same genre?

To answer these questions, we conducted an in-depth analysis for the Italian language based on automatically dependency parsed corpora aimed at assessing i) the distribution of simple and complex nominal lexicon in different genres and different language varieties for the same genre ii) the syntactic role bears by "simple" and "complex" nouns characterizing each corpus iii) the correlation between "simple" and "complex" nouns with features of complexity underlying the syntactic structure in which they occur.

In what follows we first describe the corpora considered in this study. We then illustrate how lexical and syntactic complexity have been formalized. In Section 4 we discuss some preliminary findings obtained from the comparative investigation across corpora.

2 The Corpora

Four genres were considered in this study: Journalism, Scientific prose, Educational writing and Narrative. For each genre, we chose two corpora, selected to be representative of a complex and of a simple language variety for that genre. The level of complexity was established according to the expected target audience.

The Journalistic corpora are *Repubblica* (Rep) for the complex variety, and *Due Parole* (2Par) for the simple one. Rep is a corpus of 232,908 tokens and it is made of all articles published between 2000 and 2005 on the newspaper of the same name; 2Par contains 322 articles taken from the easy-to-read magazine *Due Parole*[1], for a total of about 73K tokens.

The corpora representative of Scientific writing are *Scientific articles* (ScientArt) for the complex language variety, and *Wikipedia articles* (WikiArt) for the simple one. The former is made of 84 documents (471,969 tokens) covering various topics on scientific literature. The latter is made of 293 documents (about 205K tokens) extracted from the Italian web portal "Ecology and Environment" of Wikipedia.

For the Educational writing corpora we relied on two collections of school textbooks: the 'complex' one (EduAdu) contains 70 texts (48,103 tokens) targeting high school students, the 'simple' one (EduChi) a sample of 127 texts (48,036 tokens) targeting primary school students.

Finally, the Narrative corpora are composed by the original versions of *Terence* and *Teacher* (TTorig), for the complex pole, and the correspondent simplified versions for the simple pole. Terence, which is named after the EU Terence Project[2], is made of 32 documents, covering short novels for children. Teacher contains 24 documents extracted from web sites dedicated to educational resources for teachers. All *Terence* and *Teacher* texts have a simpler version (TTsemp), which is the result of a manual simplification process as described by Brunato and Dell'Orletta (2017).

All corpora were automatically tagged by the part-of-speech tagger described in (Dell'Orletta, 2009) and dependency parsed by the DeSR parser described in (Attardi et al., 2009).

3 Features of Linguistic Complexity

3.1 Assessment of Lexical Complexity

For each corpus we extracted all lemmas tagged as nouns, without considering proper nouns, and we classified them as 'simple' vs 'complex' nouns. Such a distinction was established according to their frequency, which is one of the most used parameter to assess the complexity of vocabulary (see Section 1). Frequency was here computed with respect to a reference corpus, i.e. ItWac (Baroni et al., 2009), which was chosen since this is the biggest corpus available for standard Italian thus offering a reliable resource to evaluate word frequency on a large-scale. After ranking all nouns for frequency, we pruned those with a frequency value ≤ 3 and we kept the first quarter of nouns as representative of the sample of *simple* nouns and the last quarter as representative of the sample of *complex* nouns for each corpus.

[1] www.dueparole.it

[2] www.terenceproject.eu

3.2 Assessment of Syntactic Complexity

To investigate our main research questions, that is how lexical complexity affects syntactic complexity and the possible influence of genre and language variety on this relationship, we focused on a set of features automatically extracted from the sentence parse tree. These features were chosen since they are acknowledged to be predictors of phenomena of structural complexity, as demonstrated by their use in different scenarios, such as the assessment of learners' language development or the level of text readability (e.g. (Collins-Thompson, 2014; Cimino et al., 2013; Dell'Orletta et al., 2014)).

For each corpus, all the considered features were computed for all occurring nouns, for the subset of *complex* nouns and for the subset of *simple* nouns. Specifically, we focused on the following ones:

- The linear distance (in terms of tokens) separating the noun from its syntactic head (*HeadDistance* in all following Tables)
- The hierarchical distance (in terms of dependency arcs) separating the noun from the root of the tree (*RootDistance*)
- The average number of children per noun (*AvgChildren*)
- The average number of siblings per noun (*AvgSibling)*

4 Discussion

To have a first insight into the effect of genre and language variety on the interplay between lexical and syntactic complexity, we compared the main syntactic roles that nouns play in the sentence by calculating the frequency of all dependency types linking a noun to its head. This is shown in Figure 1, which reports the percentage distribution of typed dependency relationships linking a noun to its syntactic head across all corpora. For each corpus there are three columns: the first one considers data for all nouns of each corpus without any complexity label, the second one only data for the *simple* noun subset and the last one only data for the *complex* noun subset.

It can be noted that the distribution of nouns used as prepositional complements (prep) is the higher one across all corpora although with differences ranging from the lowest percentage (35.5%) in the 'easy' version of the narrative corpus (i.e. *TTsemp*) to the highest one (49.9%) in ScientArt (i.e. the complex language variety for the scientific writing genre). The syntactic role of prepositional complement is especially played by *simple* nouns compared to *complex* nouns. This is particularly evident in *ScientArt* and *Repubblica*, where the difference between *simple* and *complex* nouns occurring as prepositional complements is equal respectively to 20 and 15 percentage points. Conversely, *complex* nouns are more widely used as modifiers than *simple* nouns, especially in *Repubblica*. The percentage of nouns occurring in the subject and object position is less than 20% in all corpora. Interestingly, the higher occurrence of nominal subjects is attested in *DueParole* and *ChildEdu* (14.1 and 16, respectively). This might suggest that simpler language varieties, independently from genre, make more use of explicit subjects than implicit or pronominal ones. Besides, the likelihood of a noun to be simple or complex does not particularly affect the overall presence of nominal subjects, unless for *ScientArt* and *Rep* which both show a higher percentage of *simple* nouns in the subject position.

A deeper understanding of the relationship between lexical and syntactic complexity was provided by the investigation of the syntactic features described in Section 3.2. Table 1 shows the average value of the monitored features with respect to all nouns (All), to the subset of *complex* nouns (Comp) and to the subset of *simple* nouns (Simp) extracted from all corpora. We assessed whether the variation between these feature values was statistically significant in a three different comparative scenarios: i) between the two corpora of the same genre, ii) between the complex corpora of each different genre and ii) between the simple corpora of each different genre. Table 2 shows linguistic features varying significantly for all the considered comparisons according to the Wilcoxon rank-sum test, a non parametric statistical test for two independent samples (Wild, 1997).

If we compare the two language varieties within each genre, it can be seen, for instance, that nouns are hierarchically more distant from the root in the complex than in the simple version. Such a variation, which is highly significant for all genres, affects more the Journalistic genre (*DuePa-*

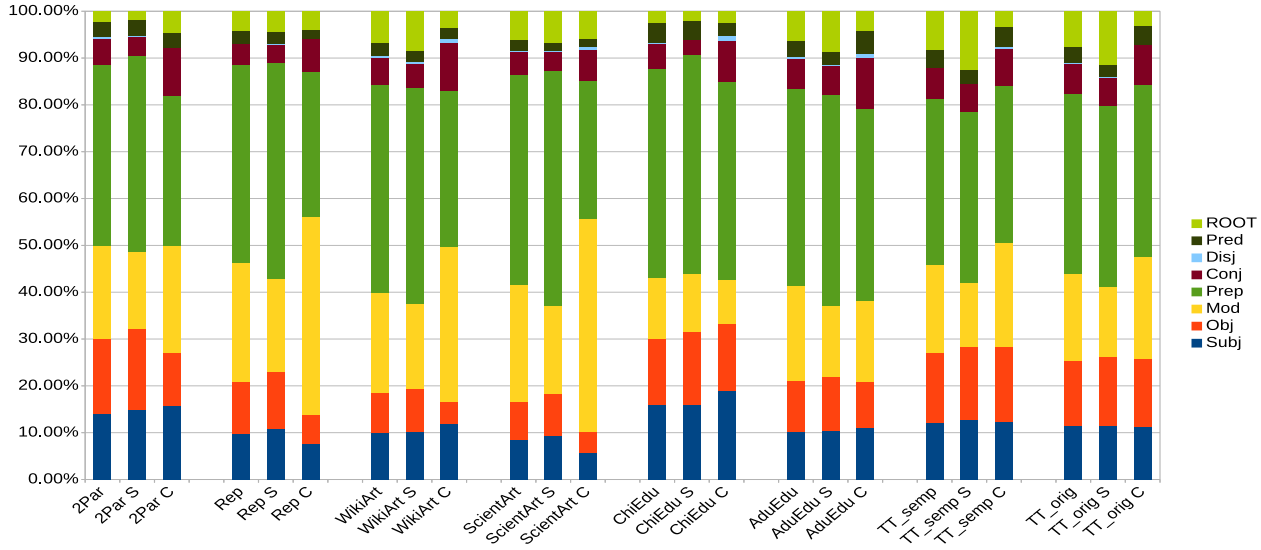

Figure 1: Distribution of typed syntactic dependencies linking nouns to their head across corpora. For each corpus, the first column refers to all nouns; the second one to the subset of *simple* nouns; the third one to the subset of *complex* nouns

	HeadDistance			**AvgChildren**			**AvgSibling**			**RootDistance**		
	All	Comp	Simp	All	Comp	Simp	All	Comp	Simp	All	Comp	Simp
2Par	2.252	2.342	2.256	1.318	1.218	1.345	1.675	1.956	1.580	2.969	2.816	2.993
Rep	2.210	2.271	2.272	1.213	0.979	1.323	1.558	1.509	1.564	4.197	4.314	4.131
Wiki	2.531	2.686	2.625	1.363	1.138	1.528	1.603	1.897	1.592	4.284	4.346	4.097
ArtScient	2.162	2.391	2.409	1.229	1.066	1.388	1.399	1.487	1.418	4.835	5.132	4.598
EduChi	2.177	2.338	2.171	1.311	1.303	1.353	1.523	1.621	1.458	3.408	3.387	3.388
EduAdu	2.598	2.875	2.695	1.440	1.375	1.560	1.654	1.715	1.640	4.269	4.483	4.143
TTsemp	2.167	2.334	2.172	1.342	1.335	1.470	1.690	1.789	1.659	3.017	2.953	2.882
TTorig	2.252	2.399	2.269	1.339	1.333	1.439	1.681	1.705	1.697	3.268	3.200	3.169

Table 1: Average value of the monitored syntactic features with respect to all nouns (All), to the subset of complex nouns (Comp) and to the subset of simple nouns (Simp) extracted from all the examined corpora.

role: 2.969; *Rep*: 4.197) and, to a lesser extent, the Educational one (*EduChi*: 3.408; *EduAdu*: 4.269). However, for the other monitored syntactic features, the *Wiki* corpus appears as slightly more difficult than its complex counterpart: it has nouns that are less close to their head (*Wiki*: 2.531; *ArtScient*: 2.162) and have a richer structure in terms of number of children (*Wiki*: 1.363; *ArtScient*: 1.229). With the exception of *root distance*, variations concerning other features within the Narrative genre are not statistically significant. This can be possibly due to the particular composition of the two selected corpora: indeed, both *Terence* and *Teacher* texts in their original version were already conceived for an audience of children and young students, and they were not greatly modified in their simplified version.

We finally assessed whether the variation of these features was statistically significant comparing the *simple* and the *complex* noun subset of the same corpus (Table 3). According to this dimension, we can observe that *complex* nouns have, on average, less dependents (*AvgChildren* feature) than *simple* ones, independently from the internal distinction within genre; on the contrary, they tend to occur more distant from the root, especially in the complex variety of Scientific prose (*ArtScient_Comp*: 5.132; *ArtScient_Simp*: 4.598).

5 Conclusion

While language complexity is a central topic in linguistic and computational linguistics research, it is typically addressed from a local perspective, where each subdomain is investigated in isola-

	HeadDistance			AvgChildren			AvgSibling			RootDistance		
	All	C	S	All	C	S	All	C	S	All	C	S
2Par vs Rep	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓	✓*	✓*	✓*
Wiki vs ArtScient	✓*	✓*	✓*	✓*	✗	✓*	✓*	✓*	✓*	✓*	✓*	✓*
EduChild vs EduAdu	✗	✗	✗	✓*	✗	✓	✓	✗	✗	✓*	✓*	✓*
TTsempl vs TTorig	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓*	✓	✓*
ArtScient vs EduAdu	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*
Rep vs ArtScient	✓*	✗	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*
Rep vs EduAdu	✓*	✓*	✓*	✓*	✓*	✓*	✓	✓	✗	✓	✗	✗
Rep vs TTorig	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*
TTorig vs ArtScient	✓*	✓*	✓*	✓*	✓*	✓	✓*	✓*	✓*	✓*	✓*	✓*
TTorig vs EduAdu	✗	✗	✗	✓*	✗	✓	✓*	✗	✓*	✓*	✓*	✓*
2Par vs EduChild	✓	✓*	✗	✓*	✓*	✗	✓*	✗	✓	✓*	✓*	✓*
2Par vs TTsemp	✗	✓*	✗	✓*	✓*	✗	✓	✗	✓*	✗	✗	✓*
2Par vs Wiki	✓*	✗	✓*	✓*	✓	✓*	✓*	✗	✓*	✓*	✓*	✓*
TTsemp vs EduChild	✗	✗	✗	✗	✗	✗	✓*	✗	✓*	✓*	✓*	✓*
TTsemp vs Wiki	✓*	✓*	✓*	✗	✓*	✗	✓*	✗	✓*	✓*	✓*	✓*
Wiki vs EduChild	✓*	✓*	✓*	✗	✓*	✓	✗	✗	✗	✓*	✓*	✓*

Table 2: Syntactic features that vary in a statistically significant way between the simple and the complex corpus of the same genre, between the complex corpora of each genre and between the simple corpora of each genre. "✗" means a non significant variation; "✓" means a significant variation at <0.05; "✓*" means a very significant variation at <0.01. All=all nouns; C=complex nouns; S=simple nouns.

	HeadDistance	AvgChildren	AvgSibling	RootDistance
2ParSostS vs 2ParSostC	✓*	✓*	✓*	✓*
RepSostS vs RepSostC	✓*	✓*	✗	✓*
WikiSostS vs WikiSostC	✗	✓*	✓*	✓*
ArtScientSostS vs ArtScientSostC	✓*	✓*	✗	✓*
EduChildSostS vs EduChildSostC	✗	✓	✗	✗
EduAduSostS vs EduAduSostC	✓	✓*	✗	✓*
TTsempSostS vs TTsempSostC	✓*	✗	✗	✗
TTorigSostS vs TTorigSostC	✓*	✓	✗	✗

Table 3: Linguistic features that vary in a statistically significant way between the *simple* and the *complex* nouns of the same corpus. "✗" means a non significant variation; "✓" means a significant variation at <0.05; "✓*" means a very significant variation at <0.01. All=all nouns; C=complex nouns; S=simple nouns.

tion. In this preliminary work, we have defined a method to study the interplay between lexical and syntactic complexity restricted to the nominal domain. We modeled the two notions in terms of frequency, with respect to lexical complexity, and of a set of parse tree features formalizing phenomena of syntactic complexity. Our approach was tested on corpora selected to be representative of different genres and different levels of complexity within each genre, in order to investigate whether noun complexity differently affects syntactic complexity according to the two dimensions. We observed e.g. that nouns tend to appear closer to the root in simple language varieties, independently from genre, while the effect of genre and linguistic complexity is less sharp with respect to the other considered features.

To have a deeper understanding of the observed tendencies we are currently carrying out a more in depth analysis focusing on fine-grained features of syntactic complexity, such as the depth of the nominal subtree. Further, we would like to enlarge this approach to test other constituents of the sentence, such as the verb.

Acknowledgments

The work presented in this paper was partially supported by the 2–year project (2017-2019) PERFORMA – Personalizzazione di pERcorsi FORMativi Avanzati, funded by Regione Toscana (Progetti Congiunti di Alta Formazione – POR FSE 2014-2020 Asse A – Occupazione) in collaboration with Meta srl company.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. In Language Resources and Evaluation, 43:3, pp. 209-226.

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

Cristina Burani. 2006. Morfologia: i processi. In: A. Laudanna and M. Voghera (cur.) *Il linguaggio. Strutture. Strutture linguistiche e processi cognitivi.* Bari, Laterza, 2006.

Isabella Chiari and Tullio De Mauro. 2014. The New Basic Vocabulary of Italian as a linguistic resource. *Proceedings of the First Italian Conference on Computational Linguistics (CLIC-IT)*, Pisa 15-19 dicembre 2014.

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistic Profiling based on General–purpose Features and Native Language Identification. *Proceedings of Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, June 13, pp. 207-215.

Kevyn Collins-Thompson. 2014. Computational Assessment of text readability. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company, 97-135.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.

Holger Diessel. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics*, 43 (3): 449–470.

Richard Futrell, Kyle Mahowald and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 91–100.

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

Edward Gibson. 2000. The dependency Locality Theory: A distance–based theory of linguistic complexity. *Image, Language and Brain*, In W.O.A. Marants and Y. Miyashita (Eds.), Cambridge, MA: MIT Press, pp. 95–126.

Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286310.

John A. Hawkins 1994. A performance theory of order and constituency. *Cambridge studies in Linguistics*, Cambridge University Press, 73.

Felice Dell'Orletta, Martjin Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. 2014. Assessing the Readability of Sentences: Which Corpora and Features. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, Baltimore, Maryland, USA.

Matti Miestamo. 2008. Grammatical complexity in a crosslinguistic perspective. In: Miestamo M, Sinnemäki K. and Karlsson F. (eds), *Language Complexity: Typology, Contact Change*, Amsterdam: Benjamins, 23–41.

Randall James Ryder and Wayne H. Slater. 1988. The relationship between word frequency and word knowledge. *The Journal of Educational Research*, 81(5):312–317.

Kortmann Berndt and Szmrecsanyi Benedikt. 2012. *Linguistic Complexity. Second Language Acquisition, Indigenization, Contact*. Berlin, Boston: De Gruyter.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59, 135–157.

Chris Wild 1997. The Wilcoxon Rank-Sum Test. *University of Auckland, Department of Statistics*