Towards an NLP-based approach for measuring syntactic complexity: preliminary experiments with Italian texts from different registers

Felice Dell'Orletta, Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" – (CNR, Italy)

The measure of syntactic complexity represents an important research topic in a variety of language-related research areas, such as language-internal variation, child language acquisition, language impairment and second language acquisition. In spite of the centrality of such a notion in different research areas, there is no single agreed-upon method or formula for the assessment of syntactic complexity. A number of measures have been proposed in the literature, focusing on a wide typology of features ranging from sentence length to the typology of grammatical constructions, to the "processing" demand, to general complexity measures based e.g. on the mean number of clauses per sentence or the occurrence of particularly complex syntactic constructions (for example, coordination, passives, embedded and/or subordinate structures).

As pointed out in [1], apart from sentence length which can be easily identified against raw text, the other features make different pre-processing demands: some of them can be recognized in the morpho-syntactically annotated text. However, more complex structures require information that goes beyond what morpho-syntactic taggers can provide: examples of such structures include the depth of the syntactic tree, the count of main clauses vs subordinate clauses, the count of different types of conjoined structures (involving, e.g. verbs, nouns, adjectives, prepositional complements) to mention only a few. This entails that the identification of most syntactic complexity features requires manual analysis of large language samples, an extremely laborious task to be carried out by skilled analysts. Availability of morpho-syntactically and/or syntactically annotated corpora can speed up the feature recognition task, but the development of linguistically annotated corpora is also time-consuming. In both cases, this has limited the size of the language samples analyzed in previous syntactic complexity research, with a potentially negative impact on the reliability of achieved results.

This suggests that there is a clear need for computational methods and techniques that can help automating the process of measuring syntactic complexity on the basis of parameters concerning the linguistic structure with sufficient accuracy. Over the last five years, there have been a few attempts to automate the process, differing with respect to the kinds of implemented complexity metrics as well as the levels of incorporated Natural Language Processing (NLP) capabilities. To our knowledge, most of them have been carried out in the framework of phycholinguistic and language acquisition research, e.g. to measure readability of texts [2,3], to assess cognitive impairment [4], to track language development in children [5] as well as in second language learners [6].

In this paper, we explore how NLP can be used to automatically identify relevant syntactic complexity features in texts with the aim of assessing their correlation with specific linguistic registers. Our final goal is twofold. On the one hand, in line with the recent attempts mentioned above, we demonstrate that automatic morpho-syntactic and syntactic annotation of texts provides sufficiently accurate output for use in the automatic extraction and measurement of syntactic complexity features. On the other hand, we identify the set of syntactic features strongly correlating with considered linguistic registers.

This exploratory study is carried out on a set of corpora of Italian texts representative of different registers, detailed in Table 1 (overleaf), by using state-of-the-art NLP technologies for Italian, namely a morpho-syntactic tagger [7] and a dependency parser [8]. Statistical language modelling techniques are employed to identify the feature sets strongly correlating with individual linguistic registers. The study is based on a wide range of linguistic features covering different aspects of the morpho-syntactic and dependency structure of texts: to mention only a few, the distribution of morpho-syntactic tags and dependency types, general complexity measures concerning the average depth of the dependency tree, the average length of dependency links, the depth of dependency "chains" (e.g. chains of prepositional complements), to more specific measures aimed at detecting complex syntactic constructions (e.g. the average number of clauses per sentence, the proportion of main vs subordinate clauses, the count of different types of conjoined structures involving verbs, nouns, adjectives, prepositional complements).

References

[1] Szmrecsányi, B. (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, March 10–12, 2004,* Presses universitaires de Louvain, Louvain-la-Neuve, pp. 1032–1039.

[2] Heilman, M., K. Collins-Thompson & M. Eskenazi (2008) An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus, Ohio: Association for Computational Linguistics.

[3] Schwarm, S. & M. Ostendorf (2005) Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 523–530.

[4] Roark, B., M. Mitchell & K. Hollingshead (2007) Syntactic complexity measures for detecting Mild Cognitive Impairment. In *Proceedings of the ACL 2007 Workshop on Biomedical Natural Language Processing (BioNLP),* pages 1-8, Prague, Czech Republic.

[5] Sagae, K., A. Lavie & B. MacWhinney (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-05).* Ann Arbor, MI.

[6] Schwarm, S.E. & M. Ostendorf (2005) Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-05).* Ann Arbor, MI. pp. 523–530.

[7] Dell'Orletta, F. (2009) Ensemble system for Part-of-Speech tagging. In Proceedings of Evalita'09, Reggio Emilia, December 2009.

[8] Attardi, G. & F. Dell'Orletta (2009) Reverse Revision and Linear Tree Combination for Dependency Parsing. In *Proceedings of NAACL HLT 2009*.

Tables

Table 1

Corpus	**Type**	**Source**	**Tokens**	**Sentences**
Repubblica 2002-2005	Newspapers	Corpus di Lingua Italiana Contemporanea (CLIC – ILC)	10.325.590	484.296
Libri 1974-1989	Fiction	Corpus di Lingua Italiana Contemporanea (CLIC – ILC)	1.243.560	58.393
Blog 2009	Blogs	Italian national FIRB project "Paisà"	130.266	5123
Corpus legislativo ambientale 1997-2005	Legal corpus	Venturi (2006)	1.645.286	74.253
Racconti fantastici 2007-2008	Imaginative prose	Marinelli *et al*. (2008)	25.492	1.158
School books 2010	Primary school textbooks	"Migrazioni" Project (CNR national project)	52.768	2.805
Due Parole 2001-2006	Easy reading magazine for people with cognitive impairment	http://www.dueparole.it/	248.304	9.393