Is this Sentence Difficult? Do you Agree?

Dominique Brunato[◇], **Lorenzo De Mattei**[•]
Felice Dell'Orletta[◇], **Benedetta Iavarone**[⋆], **Giulia Venturi**[◇]
[•]Dipartimento di Informatica, Università di Pisa
[⋆]Scuola Normale Superiore, Pisa
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato,felice.dellorletta,giulia.venturi}@ilc.cnr.it
lorenzo.dematteiadi.unipi.it beneiavarone@gmail.com

Abstract

In this paper, we present a crowdsourcing-based approach to model the human perception of sentence complexity. We collect a large corpus of sentences rated with judgments of complexity for two typologically-different languages, Italian and English. We test our approach in two experimental scenarios aimed to investigate the contribution of a wide set of lexical, morpho-syntactic and syntactic phenomena in predicting i) the degree of agreement among annotators independently from the assigned judgment and ii) the perception of sentence complexity.

1 Introduction

Linguistic complexity is a well-studied and multifaceted notion for which several measures have been proposed in different frameworks ranging from First and Second Language Acquisition, language typology and readability assessment. Such measures depend on the perspective from which linguistic complexity is considered. According to one established distinction, linguistic complexity should be divided into an *absolute* vs a *relative* notion (Miestamo, 2008). While the former is driven by theory and aims at assessing the complexity of a language according to some formal properties of the linguistic system, the latter defines complexity in relation to the language user (e.g. speaker, listener or learner) thus considering complexity in terms of processing difficulty. From this second perspective, sentence complexity is analyzed in terms of cognitive load, which can be inferred using both off-line (e.g. complexity judgments, error rates on comprehension test, preference for a structure over a meaning-equivalent one in elicited production tasks) and online processing measures (e.g. eye-tracking data such as total gaze time, fixation duration and pupil dilation). To operationalize factors underlying sentence processing performance, several complexity metrics have been proposed which consider properties of single word and sentence, as well as experience-based expectations. Word-level predictors shown to correlate with greater processing difficulties are e.g. word frequency, age of acquisition, root frequency effect, orthographic neighbourhood frequency. At syntactic level, a well-studied measure of sentence complexity takes into account dependency length (Gibson, 1998, 2000), which has been used to explain 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 (Gordon et al., 2001; Staub et al., 2010), as well as variations in word order patterns (Gildea and Temperley, 2010), also in a diachronic perspective (Gulordava and Merlo, 2015). Alternatively, processing difficulty has been explained in terms of surprisal (Hale, 2001). Computational models to calculate lexical and syntactic surprisal have been developed by e.g. Roark et al. (2009) using a broad-coverage probabilistic PCFG parser and Demberg and Keller (2009), who introduced Prediction Theory, which aims at unifying Dependency Length Theory with syntactic surprisal, by making use of a psycholinguistically-motivated version of tree-adjoining grammar.

Unlike more conventional studies on human sentence processing carried out in experimental settings, in this study we rely on crowdsourcing methods to investigate how people perceive sentence complexity. The reliability of crowdsourced data for linguistics and computational linguistics research is well acknowledged as shown in the survey by Munro et al. (2010) proving that the quality of findings obtained from the crowd is comparable, if not higher, to controlled laboratory experiments. In addition, crowdsourcing reaches a

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2690–2699
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

broader population, in terms of age, education, profession etc. and it is thus more suitable to catch the "layman" intuition of sentence complexity. For these reasons, this method has been used in recent works in the field of readability and text simplification; it is the case of Lasecki et al. (2015); Clercq et al. (2013); Brunato et al. (2016) where the crowd was asked to evaluate the level of complexity or the degree of informativeness of simplified sentences compared to the original one.

In our study, we adopted a similar perspective relying on a crowdsourcing approach to collect a wide resource containing multiple annotations of sentence complexity given by humans. Unlike traditional studies which typically assess either lexical or structural complexity phenomena, we focused on the analysis of a wide set of linguistic features to investigate how all contribute to human perception of sentence complexity. This choice is also motivated by previous studies focused on the "form" of a text all related to the assessment of complexity, e.g. readability assessment (Collins-Thompson, 2015), first language acquisition (Sagae et al., 2005) and Native Language Identification (Malmasi et al., 2017).

2 Our Contributions

Our contribution to the study of sentence complexity is multiple:

- we address two research questions aimed to investigate the role played by a set of linguistic phenomena in characterizing a) the agreement among annotators when they rated the sentences independently from the assigned score and b) the human perception of complexity.
- we introduce a new crowdsourcing-based method to assess how people perceive sentence complexity and we test it for two languages;
- we collect two corpora of sentences annotated by humans with a judgment of complexity;

The two research questions refer to two phenomena that are by definition highly subjective and difficult to define. Our study intends to address this vagueness providing the following main contributions: i) detecting the main linguistic phenomena involved in the prediction of agreement and ii) which phenomena characterize a sentence that is perceived complex by a high number of human subjects.

All the data discussed here are made available at www.italianlp.it/resources/.

3 Approach

We collected a dataset of rated sentences through a crowdsourcing task in which annotators were asked to give a score of complexity to a sentence. The task was carried out on two languages, Italian and English, which have different morpho-syntactic and syntactic properties such as morphological richness and word order freedom. This choice was aimed to investigate whether there are linguistic complexity parameters shared by typologically different languages. Starting from the collected rated sentences, we automatically extracted a wide set of features spanning across multiple levels of linguistic description, which have been acknowledged in the literature on human sentence processing to be involved in sentence complexity. The contribution of these features in modeling the perception of sentence complexity was tested in two different scenarios: i) a classification experiment to assess which features contribute more in the automatic prediction of the degree of agreement among annotators and which features vary in a statistically significant way between agreed and not-agreed sentences; ii) a regression experiment to evaluate if the considered features allow predicting the complexity judgment assigned by humans and how they contribute to the prediction.

In what follows, we introduce the three main ingredients of our approach, i.e. the set of linguistic features (Section 3.1), the datasets of sentences (Section 3.2) and the crowdsourcing task (Section 3.3). In the rest of the paper, we describe the experimental scenarios raised by our two research questions and discuss the results (Sections 4 and 5).

3.1 Linguistic Features

The set of features considered in this study captures different aspects of sentence complexity.

Raw text features:

word length, i.e. average number of characters per words (*char_tok* in all tables and figures that follow) and **sentence length**, i.e. average number

of words per sentence (*n_tokens*), which are typically used as a proxy of lexical and syntactic complexity in traditional readability metrics (Collins-Thompson, 2015);

Morpho-syntactic features:

distribution of part-of-speech types; **type/token ratio**, calculated as the ratio between the number of lexical types, the number of tokens, in terms of both lemma and forms (*ttr_form*, *ttr_lemma*); **verbal features**, i.e. the distribution of verbs according to mood (*verbs_mood*), tense (*verbs_tense*) and persons (*verbs_num_per*), and **lexical density** (*lex_density*), calculated as the ratio of content words (verbs, nouns, adjectives and adverbs) to the total tokens in a text. Psycholinguistic studies highlight that higher lexical density implies greater cognitive load (Gibson, 1993);

Syntactic features:

probability of syntactic dependency types e.g. subject, direct object, modifier, etc., calculated as the *distribution* of each type out of the total dependency types. Some syntactic relations have been shown to be harder to process, e.g. object-relative clauses and prepositional-phrase attachments (Gibson and Pearlmutter, 1994; Gibson, 2000), or the subject and object relations especially in free word-order languages;

distribution of verbal roots, i.e. the *distribution of verbal roots* out of the total of sentence roots. A lower percentage of verbal roots implies a higher number of nominal sentences which have a less-standard structure due to verb ellipsis thus possibly causing processing ambiguity;

parse tree depth features: the *depth of the whole parse tree* (*max_depth*), calculated in terms of the longest path from the root of the dependency tree to some leaf; the *depth of embedded complement chains* governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers, calculated as the *total* number of prepositional chains (*n_prep_chains*) and the *average* depth of chains (*prep_chain_l*); the *distribution of embedded complement chains by depth*, calculated as the number of chains out of the total number of chains in a sentence (*prep_depth*). All these features are related to length factors and correlate with processing difficulty (Frazier, 1985), as in the case of long sequences of embedded prepositional complements;

verbal predicate features: the distribution of verbal head (*verb_head*); the *arity of verbs*, meant as the average number of instantiated dependency links sharing the same verbal head covering both arguments and modifiers *verb_arity*); the *distribution of verbal head by arity*, calculated as the total number of verbal heads with the same arity in a sentence (*verb_head_arity*); the *relative ordering of subject and object with respect to the verbal head* (*order_subj* and *order_obj*);

subordination features include the *distribution of main vs. subordinate clauses* (*n_subord_clauses* and *n_princ_clauses*; the *average depth of chains of embedded subordinate clauses*, calculated as the *total* number of subordinate chains (*n_subord_chain*) and the *average* depth of subordinate chains (*subord_chain_l*); the *distribution of embedded subordinate clauses chains by depth*, calculated as the number of chains out of the total number of chains in a sentence (*subord_depth*). We also calculated the order of the subordinate clause with respect to the main clause (*order_subord*), since according to e.g. (Miller and Weinert, 1998), sentences containing subordinate clauses in postverbal than in preverbal position are easier to process;

length of dependency links calculated as the number of words between the syntactic head and the dependent: the feature includes the *length of all dependency links* (*links_len)* and of the *maximum dependency links* (*max_links_l*). It is widely known that long-distance constructions cause cognitive load (Gibson, 1998; Gildea and Temperley, 2010);

clause length measured as the number of tokens occurring within a clause (*token_clause*). Syntactic metrics relying on this feature, such as the T-Unit (Hunt, 1966), are widely used e.g. in first and second language acquisition to assess the development of syntactic competence.

3.2 Data

The experiments were carried out on a subset of sentences extracted from two manually revised treebanks. We chose this kind of data in order to prevent possible errors produced by the automatic annotation of sentences. Specifically, we considered the newspaper section of the Italian Universal Dependency Treebank (UDT) (Simi et al., 2014) and the automatically converted Wall Street Journal section of the Penn Treebank (McDonald et al., 2013). Since we wanted to investigate the human

perception of complexity with respect to standard language, we didn't use the English version of the UDT containing different genres of web media (e.g. blogs, emails). Although the two selected treebanks have different annotation schemes, the annotation scheme of the UDT project (McDonald et al., 2013) is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2006). This allowed us to compare linguistic phenomena correlated with sentence complexity minimizing possible cross-linguistic differences due to not uniform principles of sentence structure representation. In order to reduce the influence of lexicon on the study of sentence complexity we pruned from the two treebanks those sentences containing low-frequency lemmas with respect to a lemma frequency list that we automatically extracted from a large reference corpus, excluding numerals and proper nouns. For what concerns Italian, we used as a reference corpus PAISÁ (Lyding et al., 2014), which is one of the biggest corpus of authentic contemporary Italian texts. For English, we selected a large corpus of sentences from the Wall Street Journal (Nivre et al., 2007). For both languages, all the sentences contained in the two treebanks were grouped into 6 bins based on a different sentence length, i.e. 10, 15, 20, 25, 30, 35 tokens (only for Italian with a range of +/- 1 tokens each). This was meant to investigate if some linguistic features that are known to correlate with sentence length (e.g. parse tree depth features and dependency links) still play an influence on sentence complexity judgments when sentence length is controlled. Sentences in each subset were then ranked according to the sum of the average frequency of their lemmas. We extracted for each bin the first 200-top ranked sentences, with the exception of Italian for which the last bin contains 123 sentences. As a result of the whole selection process, we obtained 1,200 sentences for English and 1,123 for Italian used for experiments.

3.3 Collection of Judgments of Complexity

To collect human complexity judgments, we administered a crowdsourcing task through the platform CrowdFlower[1]. For each language we recruited 20 native speakers who were asked to read a sentence and rate how difficult it was on a 7-point scale where 1 means "very easy" and 7 "very difficult". Sentences were randomly ordered and presented on distinct pages containing five sentences each. To improve the quality of the collected annotations we chose workers with a "high quality" level assigned by the platform on the basis of their performance in previous tasks and we set a minimum of ten seconds to complete a page. We computed the Krippendorff's alpha reliability corresponding to the number of annotators who assigned the same judgment. We obtained a reliability of 26% for Italian and 24% for English.

[1]www.crowdflower.com

4 Studying the Agreement between Human Judgments

Our first research question concerned the investigation of linguistic phenomena characterizing the agreement among annotators in assigning the same judgment of complexity to a sentence. To this end, we split the whole set of rated sentences into ten sets corresponding to the number of annotators giving a judgment of complexity within a same range, hereafter referred to *degrees of agreement*[2]. Figure 1 reports the number of sentences for each degree of agreement. For both languages, if we consider a minimum number of 10 agreeing annotators, very few sentences were discarded (∼50 for Italian and 70 for English). As the number of agreeing annotators increases, the number of sentences progressively decreases but we still have a considerable number of sentences (∼600) when 14 annotators agree.

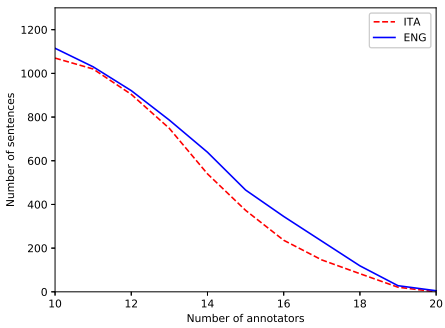


Figure 1: Number of sentences at different degrees of agreement.

To study the linguistic phenomena characterizing the agreement, we firstly extracted the features described in Section 3.1 from sentences on which annotators agreed (*agreed sentences*) and from the rest of sentences (*not-agreed sentences*); we assessed if the difference is statistically significant

[2]Each range was calculated in terms of standard deviation from the mean judgment values given to each sentence.

using Wilcoxon Rank-sum test. This was done for each agreement threshold.

We then performed a feature selection process to identify the features that maximize the accuracies of a classifier in predicting *agreed* vs *not-agreed* sentences. To create a ranking of feature relevance, we used the Recursive Feature Elimination (RFE) algorithm implemented in the Scikit-learn library (Pedregosa et al., 2011), using Linear SVM as estimator algorithm, and we dropped 1 feature in each iteration. We evaluated the classifier performance using a 3-fold cross validation method. At the end of this process we selected the top ranked features. This procedure was iterated 10 times for each degree of agreement.

In order to evaluate the accuracy of the SVM classifier we computed a baseline corresponding to the performance of the classifier using a most-likely class classification method, where each sentence is always classified into the most likely class.

Table 1 reports the features that vary in a statistically significant way (✓in table) and the ones selected in classification (marked with $\star$) for both languages and degrees of agreement levels. As it can be seen, there is an opposite trend between the statistically significant features and those selected by the classifier as the degree of agreement increases. For what concerns the Wilcoxon test, very few features have significantly different values at lower degrees of agreement. That is to say that very few features are involved in discriminating the *agreed* vs *not-agreed* sentences, especially when the agreement is lower than 14.

For both languages, raw text features (*n_tokens* and *char_tok*) vary significantly at all degrees of agreement. Interestingly, these two features are not considered by the classifier which uses more complex syntactic features, such as features related to subordination (e.g. *subord_depth*) and nominal modification (e.g. *prep_chain_l*). Syntactic features start to vary significantly as the agreement increases, e.g. parse tree depth features such as the depth of the whole parse tree (*max_depth*) and the complement chains (*dep_mark*), and features related to the use of subordination. Comparing the two languages we also found a number of differences. For example, at the lowest agreement (degree 10), features of all types turned out to vary significantly for English, while the Italian *agreed* and *not-agreed* sentences do not vary for any features. At higher agreement, Italian *agreed* sentences are characterized by the variation of two language-specific features: the position of the object with respect to the verb head (*order_obj*) and some verbal morphological features (*verbs_num_pers*, *verbs_tense*), which also contributes to the classification only for Italian.

Table 2 reports the accuracy of SVM classifier for each degree of agreement[3] and the baseline. At lower degrees of agreement (i.e. <14) the classifier achieves lower accuracy compared to the baseline showing that the selected features do not contribute to discriminate *agreed* vs *not-agreed* sentences. Instead, these features start to have a greater impact for the classification of sentences at degrees 14, 15, 16, 17. This means that at these degrees of agreement the values of the features characterizing the *agreed* sentences are significantly different from those of the *not-agreed* sentences. In addition, even though for these sentences a very high number of features are considered statistically significant by the Wilcoxon test the classifier needs less features to assign the correct class (as shown in Table 1).

5 Correlation of Linguistic Features with Sentence Complexity

The second research question aims to model the human perception of complexity studying the correlation between the set of linguistic features extracted from sentences and the judgments of complexity assigned to each sentence. We firstly calculated the average complexity judgments for the six bins of sentences of the same length (i.e. 10, 15, 20, 25, 30, 35 tokens). As expected, long sentences were judged as more complex for both languages even though all sentences were always rated as more complex for Italian (see Figure 2).

We then calculated the Spearman's rank correlation coefficient between the values of each feature and the average judgments of complexity thus obtaining a ranking of features. The correlation was computed at two distinct degrees of agreement, i.e. 10 and 14. We chose these two thresholds since at 10 the *agreed* sentences correspond to almost all the rated sentences and at 14 the SVM classifier starts to outperform the baseline (see Table 2). Besides, at 14 we still have a quite large set of *agreed* sentences allowing a reliable statistical study of the features (see Figure 1). Only at threshold 10

[3]The accuracy was computed as the average classification score of the 10 best results of the feature selection process.

Feature	Agreement															
	10		11		12		13		14		15		16		17	
	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN
char_tok	⋆	⋆	⋆	⋆	⋆	-	-	✓	✓	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
cpos_ADJ	⋆	⋆	⋆	⋆	⋆	-	-	-	✓⋆	-	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓
cpos_ADP	⋆	⋆	⋆	⋆	⋆	-	-	⋆	-	-	-	-	✓	✓	✓	✓
cpos_ADV	⋆	-	⋆	-	⋆	-	-	-	-	-	⋆	-	⋆	-	⋆	-
cpos_AUX	⋆	-	⋆	-	⋆	-	✓	-	-	-	-	-	✓⋆	-	✓	-
cpos_CONJ	⋆	⋆	⋆	⋆	⋆	-	-	⋆	-	✓	✓	✓⋆	✓⋆	✓	✓	✓
cpos_PRON	⋆	-	⋆	-	⋆	-	-	-	✓	-	✓⋆	-	✓	-	✓	-
cpos_DET	-	⋆	-	⋆	-	-	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
cpos_NUM	-	⋆	-	✓⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
cpos_PROPN	⋆	-	⋆	-	⋆	-	-	-	✓	-	⋆	-	✓⋆	-		-
cpos_PUNCT	⋆	-	⋆	-	⋆	-	✓	-	-	-	✓⋆	-	✓	-	✓⋆	-
cpos_SCONJ	⋆	-	⋆	-	⋆	-	-	-	-	-	✓⋆	-	✓	-	✓	-
cpos_VERB	-	⋆	-	⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓
dep_acl	-	-	⋆	-	⋆	-	✓	-	✓	-	✓	-	✓	-	✓	-
dep_acl:relcl	-	-	⋆	-	⋆	-	-	-	⋆	-	✓	-	✓⋆	-	✓	-
dep_adpobj	-	⋆	-	⋆	-	-	-	⋆	-	-	-	-	-	-	-	✓
dep_advcl	⋆	-	⋆	-	⋆	-	-	-	✓	-	✓	-	✓⋆	-	✓	-
dep_amod	⋆	✓⋆	⋆	⋆	⋆	✓	-	✓⋆	✓	✓	✓	✓⋆	✓⋆	✓⋆	✓	✓⋆
dep_appos	-	⋆	-	⋆	-	-	-	-	-	⋆	-	-	-	-	-	-
dep_attr	-	⋆	-	⋆	-	-	-	-	-	-	-	✓⋆	-	✓⋆	-	✓
dep_aux	-	-	⋆	-	⋆	-	✓	-	✓	-	-	-	✓⋆	-	✓	-
dep_case	⋆	-	⋆	-	⋆	-	-	-	⋆	-	-	-	✓	-	✓	-
dep_cc	⋆	⋆	⋆	⋆	⋆	-	-	-	-	✓⋆	✓	✓⋆	✓⋆	✓⋆	✓	✓
dep_ccomp	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓	-	✓	-	✓
dep_compmod	-	⋆	-	⋆	-	-	-	-	-	✓⋆	-	⋆	-	✓⋆	-	✓⋆
dep_conj	⋆	⋆	⋆	⋆	⋆	-	-	✓⋆	-	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓	✓⋆
dep_det	-	⋆	-	⋆	-	-	-	⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
dep_dobj	⋆	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓⋆	-	✓	-
dep_mark	⋆	⋆	⋆	⋆	⋆	-	✓	⋆	✓	⋆	✓⋆	⋆	✓⋆	✓	✓	✓
dep_nmod	⋆	⋆	⋆	⋆	✓⋆	-	✓	-	✓	-	-	✓⋆	✓⋆	✓	✓	✓
dep_nsubj	-	✓⋆	-	✓⋆	-	✓	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
dep_num	-	⋆	-	⋆	-	✓	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
dep_partmod	-	⋆	-	⋆	-	-	-	-	-	-	-	✓	-	✓	-	✓
dep_poss	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓	-	✓	-	✓
dep_punct	⋆	-	⋆	-	⋆	-	✓	-	-	-	✓⋆	-	✓	-	✓	-
dep_rcmod	-	⋆	-	⋆	-	-	-	⋆	-	-	-	✓⋆	-	✓⋆	-	✓
dep_xcomp	⋆	-	⋆	-	⋆	-	-	-	-	-	-	-	✓	-	✓	-
lex_density	-	⋆	-	⋆	-	-	-	✓⋆	-	✓⋆	-	✓	-	✓⋆	-	✓⋆
links_len	-	✓⋆	⋆	⋆	✓⋆	✓	✓	✓	✓⋆	✓	✓	✓⋆	✓	✓	✓	✓
max_depth	-	⋆	⋆	⋆	✓⋆	-	✓	✓	✓	✓	✓	✓	✓	✓⋆	✓	✓
max_links_l	-	✓⋆	⋆	✓⋆	✓⋆	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
n_prep_chains	⋆	✓⋆	✓⋆	⋆	✓⋆	✓	✓	✓	✓	✓	✓	✓	✓	✓⋆	✓	✓
n_principal_clauses	-	⋆	⋆	⋆	⋆	-	✓	✓	✓	✓⋆	✓	✓	✓	✓	✓	✓
n_subord_chain	⋆	⋆	⋆	⋆	⋆	✓	✓	-	✓⋆	✓⋆	✓⋆	✓	✓⋆	✓	✓	✓
n_subord_clauses	⋆	-	⋆	-	⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-
n_tokens	-	✓⋆	✓⋆	✓⋆	✓⋆	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
order_obj	-	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓	-	✓	-
order_subj	-	-	⋆	-	⋆	-	-	-	⋆	-	-	-	✓	-	✓	-
order_subord	⋆	⋆	⋆	⋆	⋆	-	✓	✓	✓	✓⋆	✓	✓	✓	✓	✓	✓
prep_chain_l	-	⋆	⋆	⋆	⋆	-	✓	-	✓	✓	✓⋆	✓	✓⋆	✓	✓	✓
prep_depth	-	✓⋆	⋆	⋆	✓	✓	✓	✓	✓	✓⋆	✓	✓⋆	✓	✓⋆	✓	✓⋆
subord_depth	⋆	⋆	⋆	⋆	⋆	-	✓⋆	⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
token_clause	-	⋆	⋆	⋆	⋆	-	-	-	-	-	-	-	✓	✓	✓	✓
ttr_form	-	✓⋆	⋆	⋆	✓⋆	✓	✓	✓	✓	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓	✓⋆
ttr_lemma	⋆	✓⋆	⋆	✓⋆	⋆	✓	✓	✓⋆	✓	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
verb_arity	⋆	⋆	⋆	⋆	✓⋆	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
verb_head_arity	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	✓⋆	⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
verb_head	⋆	⋆	⋆	⋆	✓⋆	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
verbs_num_pers	⋆	-	⋆	-	⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-
verbs_tense	⋆	-	⋆	-	✓⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-

Table 1: Linguistic features that vary statistically (✓) and the ones selected by the SVM classifier in at least 50% of the 10 runs (⋆) for Italian and English at different degrees of agreement.

	Baseline Accuracy (%) – SVM Classifier Accuracy (%)							
	10	11	12	13	14	15	16	17
Italian	95.4-95.4	91-90.8	80.6-80.5	66.7-66	51.9-**59.1**	66.8-**68.8**	79-**80.7**	87-**87.1**
English	94-94	86.8-86.8	83.6-77.4	66.3-66.1	53.9-**60**	60.7-**71.8**	70.9-**79.3**	80.4-**84.6**

Table 2: Baseline and SVM classifier accuracy at different degrees of human agreement.

we also calculated the ranking of the features with respect to the six bins of sentences of the same length (L10, L15, L20, L25, L30, L35). Figure 3 reports the ranking of features with p <0.05. Positive numbers mean that the higher the feature value the more complex the sentence was judged (i.e. the feature ranked +1 is the top-ranked one since it is the most positively correlated). Instead, negative numbers mean that the lower the feature value, the more complex the sentence was judged (i.e. the feature ranked -1 is the highest negatively correlated). In both languages, the correlation between the top 20 ranked features and the complexity judgment is extremely high, ranging from 0.30 to 0.85 when we consider sentences at agreement 14. At the two agreement thresholds, for all lengths (columns *T10*, *T14*), they concern not only sentence length but also deep syntactic

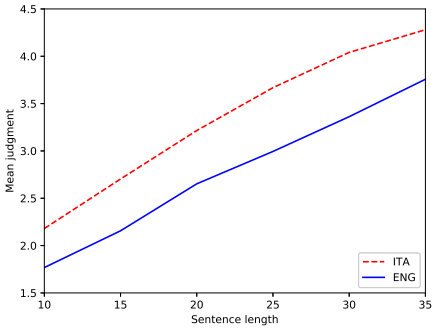


Figure 2: Mean complexity judgment at different sentence length.

features, in terms of e.g. the depth of the whole parse tree (*max_depth*), the length of dependency links (*links_len*), and features related to subordination (e.g. *n_subord_clauses*). Specifically, the 1st-ranked feature in Italian (parse tree depth) and the one in English (sentence length) have a correlation of 0.64 and 0.84 respectively. Nominal modification (*n_prep_chains*) is also highly correlated (Italian r_s=0.59, English 0.54) and similarly ranked in the two languages at 3rd position. The distribution of *verbs_num_pers* makes the sentence harder only for Italian; this is possibly related to the higher complexity of verbal morphology since the 3rd person verb in impersonal structures might increase the ambiguity of the sentence with respect to the referent. Only in English, sentence complexity is affected by the distribution of cardinal numbers (*cpos_NUM*) and the dependency type "numeric modifier" (*dep_num*), in line with the difficulty of numerical information shown in readability studies (Bautista and Saggion, 2014). Conversely, the verbal arity and the relative ordering of subjects with respect to the verb have a lower position in the negative ranking, suggesting that these features make a sentence easier: this might be due to a more fixed predicate-argument structure and word order in this language.

If we focus on sentences of the same length, features considered as a proxy of lexical complexity are in the top positions in both languages. It is the case of the average word length (*char_tok*) and the lexical density (*lex_density*) only for English. Interestingly, while for English the majority of features are similarly ranked in all bins of sentences of the same length, for Italian we observe differences between the rankings of features extracted from sentences $\leq$ and $\geq$20 token long. Namely, when the average sentence length is $\geq$20 tokens, features related to subordination make the sentence more complex.

5.1 Predicting Human Complexity Judgments

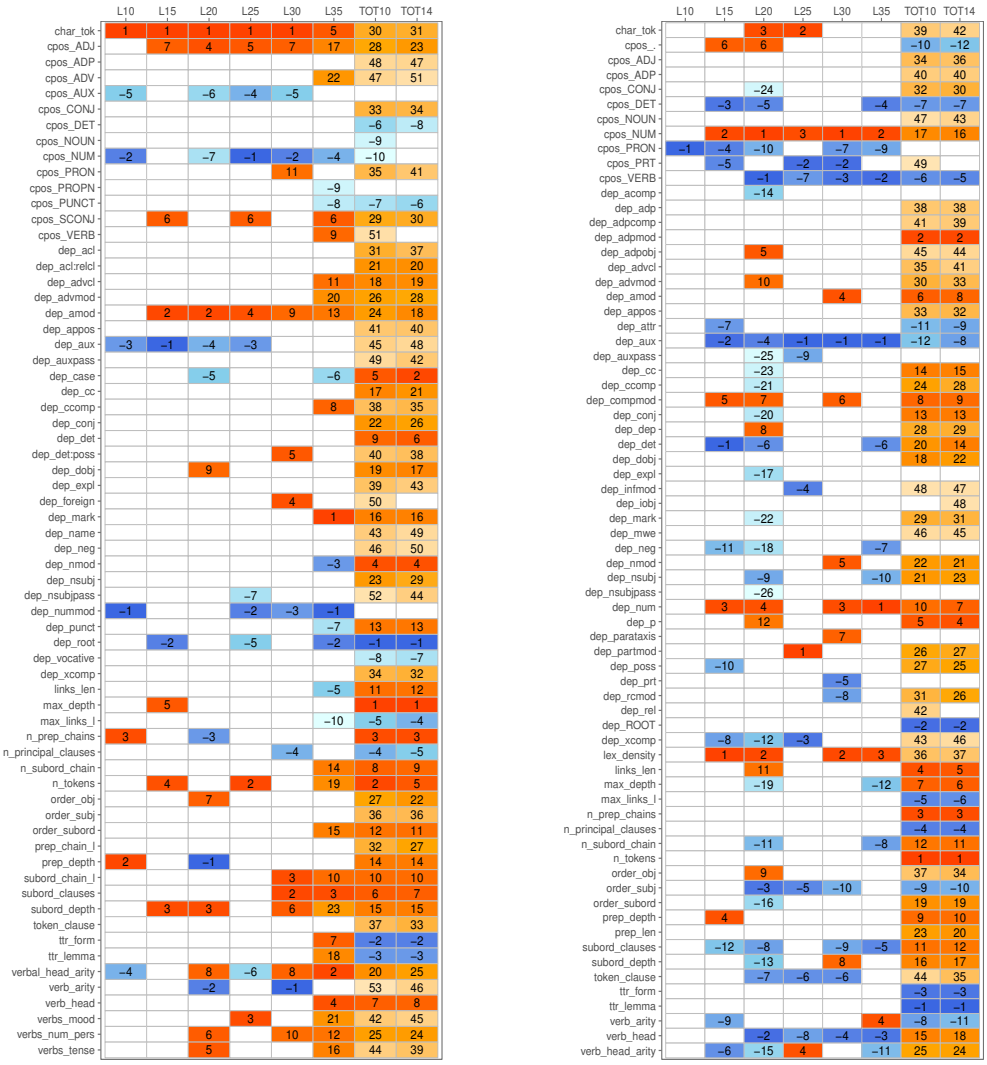
To asses the contribution of the linguistic features to predict the judgment of sentence complexity we trained a linear SVM regression model with default parameters. We performed a 3-fold cross validation over each subset of *agreed* sentences at agreement 10 and 14. We measured two performance metrics: the *mean absolute error* to evaluate the accuracy of the model to predict the same complexity judgment assigned by humans; the *Spearman correlation* to evaluate the correlation between the ranking of features produced by the regression model with the ranking produced by the human judgments. Table 3 reports the results and the average score of the two metrics. As it can be seen, the model is very accurate and achieves a very high correlation ($>$0.56 with p $<$0.001) with an average error difference (*avg mean abs err*) below 1. In particular, the model obtained higher performance in predicting the ranking of features extracted from sentences at agreement 14. This might be due to the fact these sentences are characterized by a more uniform distribution of linguistic phenomena and that these phenomena contribute to predict the same judgment of complexity. This is in line with the results obtained by the SVM classifier in predicting agreement (Table 2). This is particularly the case of English and it possibly suggests that the set of sentences similarly judged by humans are characterized by a lower variability of the values of the features.

	IT-10	**IT-14**	**EN-10**	**EN-14**
mean abs err 1	0.77	0.78	0.71	0.68
Spearman 1	0.57	0.64	0.68	0.64
mean abs err 2	0.79	0.80	0.70	0.70
Spearman 2	0.55	0.63	0.67	0.73
mean abs err 3	0.85	0.75	0.77	0.60
Spearman 3	0.55	0.64	0.61	0.71
avg mean abs err	**0.80**	**0.78**	**0.72**	**0.66**
avg Spearman	**0.56**	**0.63**	**0.65**	**0.69**

Table 3: Performance of the linear SVM regression model and the avg score at different agreements.

6 Discussion and Conclusion

In this paper, we introduced a method to model the human perception of sentence complexity relying on a new corpus of Italian and English sen-



(a) Italian. (b) English.

Figure 3: Features correlating with human judgments at different sentence lengths and with respect to the sentences at agreement 10 (*TOT 10*) and 14 (*TOT 14*).

tences rated with human complexity judgments. We tested the contribution of a wide set of linguistic features automatically extracted from these sentences in two experimental scenarios. The first one highlighted that we can reliably predict the degree of agreement between human annotators, independently from the assigned judgment of complexity: given the high subjectivity of the task, this is a quite notable result that to our knowledge has never been reported. We observed in particular that deep syntactic features related to e.g. the use of subordination and nominal modification play a main role in the automatic prediction of human agreement. This is true for the two languages even though we found that some features resulted to be more relevant in the classification of *agreed* Italian sentences, e.g. the relative ordering of the object. Interestingly, we also noticed that the classifier needs few features to predict agreed sentences when more than half of annotators shares the same judgment.

In the second experiment, we studied the correlation between linguistic features and complexity judgments. The resulting ranking highlighted the key role played by syntactic phenomena: features related to sentence structure are among the top-ranked features characterizing sentences that were rated highly complex by a given number of agreeing annotators. When sentence length was controlled, the relevance of the considered features changes in particular for Italian: e.g. features concerning the use of subordination make the sen-

tence more complex when sentence length is $\geq$20 tokens. As showed by the results of the regression model, the set of studied features contribute significantly to automatically predict the human judgment of sentence complexity.

In addition, the presented corpus can be useful for different applications. From a psycholinguistic perspective, it can be used for comparison with data collected through controlled experimental scenarios assessing sentence complexity in terms of cognitive measures (offline and online), which are also more constrained and costly to acquire in large-dimensions. The corpus also allows to study whether features of linguistic complexity are implied in modeling other properties of texts, such as the level of engagement or subjectivity. From a NLP perspective, the corpus can be exploited to train systems able to predict people's perception of complexity. For example, it can support a range of related tasks, such as the development of linguistically-informed algorithms for the automatic assessment of text difficulty, as well as in Natural Language Generation tasks, going from text simplification to the automatic generation/evaluation of highly-engaging texts.

References

Susana Bautista and Horacio Saggion. 2014. Can numerical expressions be simpler? Implementation and demostration of a numerical simplification system for spanish. In *Proceedings of LREC, the 9th International Conference on Language Resources and Evaluation*.

D. Brunato, A. Cimino, F. Dell'Orletta F., and G. Venturi. 2016. PaCCSS–IT: A parallel corpus of complex–simple sentences for automatic text simplification. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 10–18.

O. De Clercq, V. Hoste, B. Desmet, and P. Van Oosten. 2013. Using the crowd for readability prediction. *Natural Language Engineering*, pages 1–33.

Kevyn Collins-Thompson. 2015. Computational assessment of text readability: A survey of current and future research. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165(2):97–135.

Vera Demberg and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of CogSci 2009, the 31st Annual Conference of the Cognitive Science Society*, pages 1888–1893.

Lin Frazier. 1985. Syntactic complexity. *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, Cambridge: Cambridge University Press.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 24(11):1–76.

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

Edward Gibson and Neal J. Pearlmutter. 1994. A corpus-based analysis of psycholinguistic constraints on prepositional-phrase attachment. *Perspectives on sentence processing*.

Timothy R. Gibson. 1993. Towards a discourse theory of abstracts and abstracting. *Monographs in Systemic Linguistics*, Nottingham: Department of English Studies, University of Nottingham.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Peter C. Gordon, Randall Hendrick, and Marcus Johnson. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27(6):1411–1423.

Kristina Gulordava and Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of latin and ancient greek. In *Proceedings of Depling 2015, the Third International Conference on Dependency Linguistics*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the NAACL*, pages 159–166.

Kellogg W. Hunt. 1966. Recent measures in syntactic development. *Elementary English*, 43(7):732–739.

W. S. Lasecki, L. Rello, and J. P. Bigham. 2015. Measuring text simplification with the crowd. In *Proceedings of the 12th Web for All Conference (W4A 15)*.

V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, and V. Pirrelli. 2014. The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WAC-9) EACL*, pages 36–43.

S. Malmasi, E. Keelan, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, pages 449–454.

R. McDonald, J. Nivre, Y. Quirmbach-brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Tackstrom, C. Bedini, N. B. Castelló, and J. Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

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

Jim Miller and Regina Weinert. 1998. Spontaneous spoken language. *Syntax and discourse. Oxford, Clarendon Press*.

R. Munro, S. Bethard, V. Kuperman, V.T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*, pages 122–130.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Brian Roark, A. Bachrach, C. Cardenas, and C. Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top–down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 324–333.

K. Sagae, A. Lavie, and B. MacWhinne. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the ACL*.

Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is more? Towards a reduced inventory of categories for training a parser for the italian stanford dependencies. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, (LREC'14)*, pages 83–90.

A. Staub, S.J. White, D. Drieghe, E.C. Hollway, and K. Rayner. 2010. Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*.