On the order of words in Italian: a study on genre vs complexity

Dominique Brunato and **Felice Dell'Orletta**
Consiglio Nazionale delle Ricerche
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

In this paper we present a cross-genre study on word order variation in Italian based on automatically dependency–parsed corpora. A comparative analysis focused on dependency direction and dependency distance for major constituents in the sentence is carried out in order to assess the influence of both textual genre and linguistic complexity on the distribution of phenonemena of syntactic markedeness.

1 Introduction

It is almost impossible to classify languages according to a unique, universally valid, metric of complexity. However, scholars agree on a set of properties that, at different levels of linguistic description, can be viewed as "universal" parameters of complexity across languages (McWorther, 2001; Ferguson, 1982). At syntactic level, this is the case e.g. of word order freedom, i.e. the property for which the order of elements in a sentence can vary while conveying the same meaning. According to different perspectives, free–word order languages are considered as more complex than fixed–order languages.

In linguistic and psycholinguistic literature, several explanations have been given to account for word order freedom. Information structure theory assumes that the order of words in the sentence is determined by semantic and discourse pragmatic forces (Diessel, 2005); conversely, for *performance*–related accounts unmarked structures are generally preferred by the speaker because of efficiency pressures and information structure becomes relevant only if two or more alternative orders are equally difficult to process (Hawkins, 1994; Gibson, 1998; Gibson, 2000).

Also from a Natural Language Processing (NLP) perspective, it is acknowledged that parsing free–word order languages is more challenging than parsing fixed–order languages in many respects. Based on a comparative analysis of Latin and Ancient Greek treebanks, the study of Gulordava and Merlo (2015) e.g. demonstrated that word order freedom, defined as the distance between the actual dependency length of a sentence and its optimal dependency, is a source of complexity which can be inferred both from lower parsing performance and from a trend toward more fixed word orders over time. Comparing the accuracy of dependency parsing on dative alternations in English, German and Russian, Dakota et al. (2015) showed that the larger the number of possible alternative orders to parse the more training data is needed. The effect of data sparness on the automatic analysis of free word order language was also assessed in the study of Alicante et al. (2012) aimed at comparing the performance of constituency and dependency parsing on an Italian treebank.

In this paper we want to focus the attention on word order variation from a less–investigated perspective, aimed at assessing the influence of textual genre and linguistic complexity on the preservation of the basic (or unmarked) position of major constituents in the sentence, i.e. subject, object, adjective, adverb and subordinate clause. To this end, we carried out a corpus-based study for Italian – a Subject-Verb-Object (SVO) language – comparing the distribution of head–initial and head–final syntactic pairs across different textual genres and different language varieties, i.e. a "complex" one and a "simple" one for each genre, defined according to the expected target reader.

Differently from more traditional studies on word order variation in Italian e.g. (Fiorentino, 2009), this work relied on corpora automatically parsed up to the level of syntactic dependency annotation; this allowed us to carry out a broad comparative analysis of fine–grained features related

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 25-31,
Pisa, Italy, September 18-20, 2017

to word order variation according to genre and linguistic complexity, such us the average linear distance between the dependent and its head and the average depth of the syntactic tree of the dependent element, both in the canonical and non–canonical position.

2 Related Works

Syntactically annotated corpora have been promoted by several scholars as a valuable resource in the study of word order variation and related properties, especially from the perspective of language typology.

By relying on dependency direction as a typological index, Liu (2010) quantified the distribution of right- and left-branching constructions in 20 languages. Not only this study supported traditional typological classes with large quantitative data, but also provided evidence that a dominant order exists for languages left unspecified with respect to some grammatical relations (e.g. verb-object) in well-established classifications (Haspelmath et al., 2005). A similar methodology has been applied by (Liu, 2010), who conducted a comparative study based on 15 treebanks demonstrating that dependency direction is a reliable index to explain both the syntactic drift from Latin to Romance languages and to classify Romance languages as a distinct sub–group from other languages. In Futrell et al. (2015) a large cross–linguistic analysis was carried out using dependency treebanks for more than 30 languages; the comparative study allowed the authors to confirm the correlation between high order freedom and overt case–marking.

Word order variation is generally investigated together with the effect it has on dependency distance, i.e. the distance between words and their parents, typically measured in terms of intervening words. With this respect, data from dependency annotated corpora highlight that, when two or more alternative orders are possible, languages tend to prefer the order that reduces the distance between the head and its dependent (Gildea and Temperley, 2010; Futrell et al., 2015); this also holds when the examined span affects only few words, such as in the nominal domain (Gulordava et al., 2015). Such findings are thus proposed as a further demonstration that dependency length minimisation, whose effect has been widely documented in sentence processing (e.g. (Gibson, 1998; Gibson, 2000)), is a universal principle of human language.

In this paper, we want to investigate whether and to what extent word order phenomena in Italian are also influenced by textual genre. Similarly to the recent work by Liu (2017) for the English language, we focus on the two main syntactic parameters which, in a syntactic dependency paradigm, allow quantifying the effects of word order variation, i.e. dependency direction and dependency distance. The novelty of our study is that we introduce a further dimension of comparison, i.e. the level of complexity within genre, which was defined in according to the intended target reader; this was meant to assess whether some genre–specific stylistic features exist and how they affect word order properties independently from the level of complexity used in text.

In what follows, we first illustrate the corpora used in our study (Section 3) and the typology of syntactic patterns on which we focused to investigate word order variation (Section 3.1). In Section 4 we discuss the main findings of the comparative analyses carried out according to genre and linguistic complexity.

3 Data

As shown in Table 1, four genres were considered in this study: Journalism, Educational writing, Scientific prose and Narrative. For each genre, we collected two corpora, representative of a "complex" and a "simple" language variety for that genre, which were defined according to the expected readership.

The journalistic corpus is sub–divided into a corpus of general newspaper articles, *La Repubblica* (Rep), which is made of all articles published between 2000 and 2005 and a corpus of easy–to–read articles published in *Due Parole*, a monthly magazine written by linguists expert in text simplification using a controlled language for an audience of adults with a rudimentary literacy level or mild intellectual disabilities (Piemontese, 1996). The Educational corpus is articulated into two collections targeting high school (AduEdu) vs primary school (ChiEdu) students. For scientific prose, the "complex" variety is represented by a corpus of ∼470,000 tokens of scientific literature covering various topics, e.g. climate change, linguistics, while the "simple" variety is represented by a corpus of Wikipedia articles of ∼200,000 to-

Genre	Corpus	Tokens
Journalism	Repubblica (Rep)	232,908
	DueParole (2Par)	72,884
Educational	Educational materials for high-school (AduEdu)	47,805
	Educational materials for primary school (ChilEdu)	23,192
Scientific Prose	Scientific articles on specialized topics (ScientArt)	471,979
	Wikipedia articles "Ecology and Environment" portal (WikiArt)	204,460
Narrative	Terence&Teacher-original versions (TT orig)	27,833
	Terence&Teacher-simplified versions (TT simp)	25,634

Table 1: The corpora used in the study.

kens, extracted from the Italian Portal "Ecology and Environment". For what concerns the narrative texts, we relied on the resource described in Brunato et al. (2015), which was specifically developed for the study of automatic text simplification in Italian. The resource is made up of two sub–corpora, *Terence* and *Teacher*, representative of two different simplification strategies, the "structural" and the "intuitive" one respectively. Both *Terence* and *Teacher* contain two versions of the same text aligned at sentence-level, namely the authentic version of text and its manually simplified version targeting specific categories of readers. In particular, *Terence* comprises 32 short novels and their simplified version addressing hearing and deaf children, aged between 7–11, affected by text comprehension difficulties. *Teacher* is composed by 24 pairs of original–simplified texts, where the simplification was mostly carried out by a teacher for L2 students. To allow comparing the effect of linguistic complexity within this genre, we created a unique corpus of "complex" narrative texts (*TT orig*) containing only the authentic texts of both *Terence* and *Teacher* and a unique corpus of "simple" narrative texts (*TT simpl*), containing only the simplified versions.

3.1 Automatic Linguistic Analysis and Linguistic Features

All corpora selected for this study were automatically tagged by the part–of–speech tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi et al., 2009) using Support Vector Machine as learning algorithm. DeSR is trained on the ISST–TANL treebank, which mainly includes articles from newspapers and periodical, and it achieves a performance of 83.38% and 87.71% in terms of LAS and UAS when tested on matching training data. However, it is well-known that the accuracy of parsers decreases when tested against texts of a different typology from those used in training (Gildea, 2001). Thus we can assume that the performance of DeSR will probably be worse in the analysis of texts representative of e.g. narrative and scientific writing. Despite this fact, we expect that the distributions of errors will be almost similar, at least when parsing texts of the same domain and language variety, thus allowing us to carry out a reliable internal comparison with respect to the examined syntactic patterns. In addition, the effect of genre variation on the performance of a general–purpose parser is likely to be less strong since all genres here considered contain *standard* texts, i.e. texts linguistically similar to the ones used in training.

4 Data Analysis

Based on the output of the multi–level linguistic annotation, all corpora were searched for relevant syntactic features, i.e. features related to the order and linear distance between the "dependent" element and its "head" in a syntactic dependency representation.

Specifically, we focused on the following elements: subject, object, adjective, adverb and subordinate clause. For each of them we calculated *i)* the percentage distribution in the canonical and non–canonical position (i.e. the position syntactically and/or pragmatically marked), according to the predominant SVO order in Italian, and, for each position, *ii)* the linear distance (in terms of number of tokens) between the element and the relative head.[1]

[1]For what concerns the subordinate clause, the linear distance is calculated as the average number of tokens between the POS of the root of the subordinate clause sub-tree and the verb of the main clause.

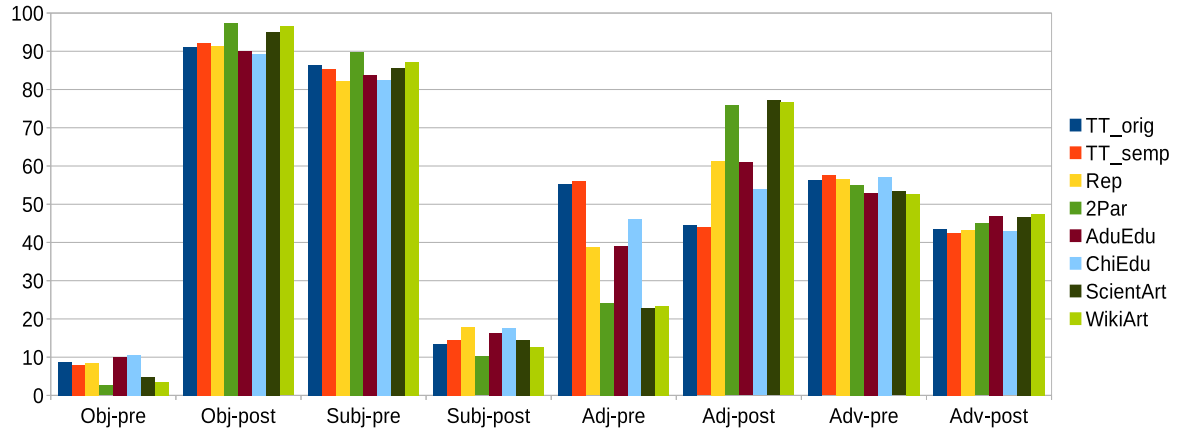

Figure 1: Percentage distribution of preverbal (Obj-pre) and postverbal objects (Obj-post), preverbal (Subj-pre) and postverbal subjects (Subj-post), prenominal (Adj-pre) and postnominal adjectives (Adj-post) and preverbal (Adv-pre) and postverbal adverbs (Adv-post) across corpora.

We also conducted a more in–depth study on subordination examining the following features: *iii)* the average length (in tokens) of the whole subordinate clause and *iv)* the average depth of the subordinate clause, calculated in terms of the longest path from the root of the subordinate sub–tree to some leaf.

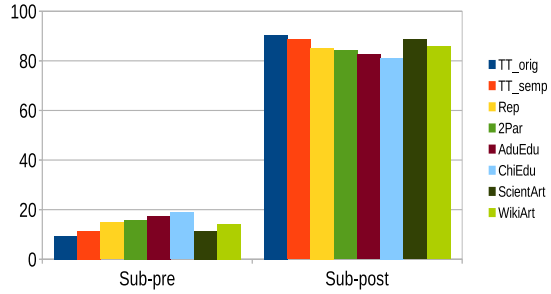

Figure 2: Percentage distribution of preverbal (Sub–pre) and postverbal subordinate clauses (Sub–post) across corpora.

Figure 1 and Figure 2 compare the percentage distribution of all the examined orders across the corpora. Let's analyse first the elements which, in all corpora, tend to occur more in their canonical position, i.e. the subject and the object.

With respect to the object, we observe that scientific texts adhere the most to the canonical order, independently from the complexity of text (ScientArt: post–verbal object: 95.14%; WikiArt: post–verbal object: 96.46%, $p<0.05$ [2]).; on the contrary, in narrative and especially in educational texts, the distribution of the unmarked object position decreases (AduEdu: 90%; ChiEdu: 89.33%, $p<0.05$). Interestingly, with the only exception of educational texts where the distribution of pre–verbal objects is almost similar in the two varieties (i.e. 9.99% vs 10.67%), all other genres show the expected positive correlation between canonical order and linguistic complexity; this is particularly evident in the journalistic genre, which reports a statistically significant difference ($p<.001$) of more than six percentage points with respect to the distribution of preverbal objects (*Rep*: 8.54%; *2Par*: 2.57%).

If scientific texts have a more rigid verb–object structure, they allow longer dependencies when the object follows the verb compared to all other genres (see the first two columns of Table 2). Such a finding is not influenced by the level of linguistic complexity within genre, since both the complex and the simple variety obtain almost equal values (~ 2.70).

As in the case of the object, also with respect to the subject, the expected correlation between the canonical SV order and the use of a simple language variety is particularly evident in the journalistic genre: indeed, texts belonging to *Due Parole* tend to preserve this order in almost 90% of cases, that is almost 7% more than their "com-

[2]Statistical significance of the difference is calculated using Mann–Whitney U test.

Corpus	Object				Subject				Adjective				Adverb			
	Pre-V		Post-V		Pre-V		Post-V		Pre-N		Post-N		Pre-V		Post-V	
	AvD	SD	AvD	SD	AvD	SD	AvD	SD	AvD	SD	AvD	SD	AvD	SD	AvD	SD
TT orig	-0.25	0.84	2.30	1.71	**-2.34**	2.24	0.57	1.67	-0.72	0.56	0.67	0.71	-1.53	2.41	0.81	1.90
TT semp	-0.21	0.8	2.25	1.58	**-2.01**	1.76	0.54	1.44	-0.73	0.58	0.63	0.66	-1.39	1.95	0.69	1.12
Rep	**-0.36**	1.43	2.56	2.22	-3.31	3.7	**0.88**	2.48	**-0.67**	0.73	0.94	0.84	-1.54	2.71	0.70	1.31
2Par	**-0.08**	0.42	2.39	1.61	-2.86	2.59	**0.51**	1.77	**-0.36**	0.61	0.96	0.60	-1.92	2.97	0.73	1.80
AduEdu	**-0.46**	1.64	**2.62**	2.20	**-3.23**	3.83	**1.09**	2.99	**-0.71**	0.65	**1.03**	1.28	-1.4	2.15	0.94	2.44
ChiEdu	**-0.26**	0.72	**2.35**	2.42	**-2.30**	2.3	**0.80**	2.17	**-0.66**	0.54	**0.91**	1.05	-1.59	2.3	0.74	1.08
ScientArt	**-0.33**	1.59	2.71	2.38	**-3.90**	4.27	**0.93**	2.86	**-0.52**	0.67	**1.12**	0.72	-0.52	0.67	0.97	2.71
WikiArt	**-0.20**	1.20	2.70	2.60	**-3.47**	3.72	**0.81**	2.67	**-0.5**	0.6	**1.1**	0.7	-1.5	2.79	0.91	2.30

Table 2: Average distance (AvD) and standard deviation (SD) of the Object, Subject, Adjective and Adverb with respect to the relative verbal (V) or nominal head (N). For values marked in bold, the difference within genre is statistically significant using Mann–Whitney U test.

.

plex" counterpart (*Rep*: 82,20%; *2Par*: 89,82%; $p<0.01$). On the contrary, both in narrative and educational texts, post–verbal subjects occur slightly more in the "simple" than in the "complex" variety, although the difference is statistically significant only for educational texts (*AduEdu*: 16.25%; *ChiEdu*: 17.61%; $p<0.05$).

For what concerns the narrative genre, it is worth noticing that the complex variety here examined is actually simpler than the complex variety of all the other genres; this is because the original texts of both *Terence* and *Teacher* are primarily written for a young readership. However, this finding should be more properly investigated in other corpora of the same genre because it might suggest that some marked constructions, such as post–verbal subjects, are genre–specific features allowing the writer to preserve the thematic progression in adjacent sentences and improve text cohesion. In this sense, such features are also maintained in the simplification process.

For what concerns educational materials, this is a heterogeneous genre comprising texts belonging in principle to different genres, ranging e.g. from fiction to scientific writing or reportage, thus making it difficult to detect the effect of language complexity.

Differently from the subject and the object, the order of adjectives within the nominal phrase is less rigid in Italian. Generally speaking, although the unmarked position of the adjective is post–nominal, it changes according to the semantic properties that the adjective carries with respect to the noun (Cinque, 2010). The relatively free ordering of adjective is confirmed by the empirical data obtained in this study, although the preferred position changes according to genres. Specifically, all but narrative genre prefer post–nominal adjectives, which is also the order that yields on average longer dependencies from the nominal head (see columns 6 and 7 in Table 2). When the internal distinction is taken into account, a stronger effect is reported by the journalistic genre, which shows a high statistically significant difference of almost 15% percentage points with respect to the distribution of post–nominal adjective (*Rep*: 61.31%; *2Par*: 75.82%; $p<0.001$).

Like the adjective, also the adverb has some degree of freedom in Italian since the unmarked position following the verb is quite flexible and in-

Corpus	Subordinate clause					
	Pre–verbal Subordinate Clause					
	AvD	SD	Length	SD	Depth	SD
TT orig	-1.27	(3.7)	1.17	(3.55)	0.51	(1.45)
TT semp	-1.1	(3.09)	1.01	(2.80)	0.50	(1.40)
Rep	-2.08	(5.60)	1.7	(4.51)	0.75	(1.83)
2Par	-1.85	(4.56)	1.4	(3.26)	0.71	(1.62)
AduEdu	-2.69	(5.72)	**2.34**	(4.96)	1.01	(2.07)
ChilEdu	-2.58	(5.36)	**2.05**	(4.19)	0.86	(1.73)
ScientArt	-2.64	(6.64)	2.15	(5.42)	1.00	(2.36)
WikiArt	-2.16	(5.60)	1.78	(4.69)	0.79	(1.91)
	Post–Verbal Subordinate Clause					
TT orig	3.01	(3.23)	**8.10**	(6.28)	3.91	(2.16)
TT semp	2.63	(2.56)	**7.04**	(4.88)	3.67	(2.19)
Rep	3.02	(3.91)	**10.33**	(9.89)	**4.49**	(3.12)
2Par	2.61	(2.51)	**7.26**	(6.70)	**3.73**	(2.47)
AduEdu	3.02	(3.68)	**11.11**	(11.04)	**4.57**	(3.32)
ChiEdu	2.63	(2.90)	**7.60**	(7.38)	**3.42**	(2.61)
ScientArt	**3.36**	(4.91)	**13.49**	(11.78)	**5.70**	(3.84)
WikiArt	**3.87**	(4.80)	**12.04**	(10.99)	**5.06**	(3.27)

Table 3: Average distance from the main clause (AvD), length and depth of the subordinate clause in the pre– and post–verbal position. For each parameter, standard deviation (SD) is reported. For values marked in bold, the difference within genre is statistically significant using Mann–Whitney U test.

fluenced by the semantic class of the adverb (Bonvino et al., 2008). The analysis across corpora shows that the predominant position is always pre–verbal; interestingly, this order is never affected by the level of complexity within each genre.

For what concerns the subordinate clause, all genres exhibit a sharp tendency to place the subordinate clause after the main clause: in a SVO language like Italian, this is the ordering that allows the parser to recognize the constituents domains more rapidly and efficiently, as predicted by performance–based theories (Hawkins, 1994). According to this parameter, narrative texts appear as the easiest ones, since the post–verbal position of the subordinate clause reaches almost 90% both in the complex and the simple variety. On the other hand, educational texts deviate more from this order, showing a higher distribution of subordinate clauses preceding the main clause (*AduEdu*: 17.38%; *ChiEdu*: 18.89%). As expected, the greater complexity derived from placing the subordinate clause before the main clause affects the internal structure of the subordinate clause at different levels (Table 3): pre–verbal subordinate clauses tend to be structurally simpler both in terms of length (i.e. they are much shorter than post–verbal ones) and depth (i.e. pre–verbal subordinate clauses have a less–embedded structure).

5 Conclusion

We have presented a study based on automatically–dependency parsed corpora aimed at quantifying the influence of textual genre and linguistic complexity on the order of constituents in Italian. On the first side, we showed that the journalistic and scientific genre tend to preserve the basic order of constituents, differently from narrative and educational texts which exhibit a higher distribution of marked orders. On the second side, the expected correspondence between the use of a simple language and the preservation of more canonical word orders has been shown to be genre–dependent: it was mainly verified within the journalistic genre, whereas narrative and educational texts tend to preserve the non–canonical order of some constituents (e.g. post–verbal subject) also in the relative "simple" variety.

Current developments of this work go in several directions: one is to conduct a thorough analysis of the impact of errors derived by automatic linguistic annotation on the distribution of the examined linguistic parameters; another is to collect corpora distinct for genre and level of linguistic complexity in other languages in order to assess whether the effect of these variables on word order variation is also language–dependent.

References

Anita Alicante, Cristina Bosco, Anna Corazza, Alberto Lavelli. 2012. A treebank–study on the influence of Italian word order on parsing performance. In *Proceedings of LREC 2012*. Istanbul, Turkey.

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.

Elisabetta Bonvino, Mara Frascarelli, Paola Pietrandrea. 2008. Semantica, sintassi e prosodia di alcune espressioni avverbiali nel parlato spontaneo. *La comunicazione parlata*, Massimo Pettorino, Antonella Giannini, Marianna Vallone, Renata Savy (Eds), Napoli, Liguori, 565–607.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of LAW IX - The 9th Linguistic Annotation Workshop*. Denver, Colorado, Giugno 2015.

Guglielmo Cinque. 2010. *The syntax of adjectives: A comparative study*. In *MIT Press*.

Daniel Dakota, Timur Gilmanov, Wen Li, Christopher Kuzma, Evgeny Kim, Noor Abo Mokh and Sandra Kübler. 2015. Do free word order languages need more treebank data? Investigating dative alternation in German, English, and Russian. In *Proceedings of the 6th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL)*, Bilbao, Spain, 14–20.

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.

Charles A. Ferguson. 1982. Simplified registers and linguistic theory. *Exceptional language and linguistics*, In Obler L.K. and L. Menn (eds.), New York, Academic Press, 49-68.

Giuliana Fiorentino. 2009. Complessità linguistica e variazione sintattica. In *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, (2), 281-312.

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. 2001. Corpus variation and parser performance. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA.

Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310.

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 the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden, August 24–26, pp. 121–130.

Kristina Gulordava, Paola Merlo and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, July 26–31, pp. 477–482.

Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie (eds.). 2005. The World Atlas of Language Structures. *Oxford: Oxford University Press*.

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

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Haitao Liu and Chunshan Xu. 2012. Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics* 48(4), 597–625.

John H. McWorther. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, 5, 125–166.

Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnodid.

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