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

Dominique Brunato, Felice Dell'Orletta

Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR),

ItaliaNLP Lab - *www.italianlp.it*

Depling 2017, Pisa, September 18-20

www.italianlp.it

Theoretical background and motivations

- Word order freedom is generally viewed as a metric of complexity at syntactic level across languages.
- According to different perspectives (typology, computational linguistics, psycholinguistics), free–word order languages are considered as more complex than fixed–order languages.
- Very few studies take into account the role of textual genre on the distribution of word order variation and related phenomena.

Our study

- A cross-genre investigation on word order variation in Italian based on automatically dependency–parsed corpora.
- For the main elements of the sentence (i.e. subject, object, adjective, adverb and subordinate clause) linguistic features related to dependency direction and dependency distance were compared across corpora.
- Comparative analysis carried out with respect to **genre** and **linguistic complexity**.

Dataset

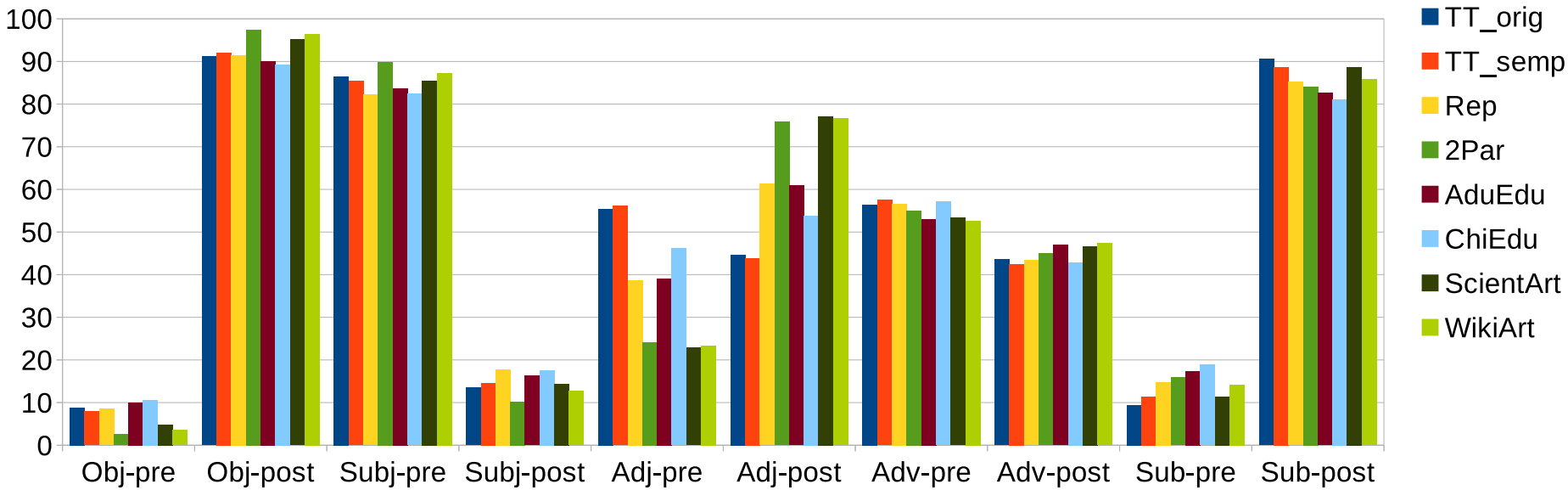
Genre	Corpus	Tokens
Journalism	Repubblica	232,908
	DueParole	72,884
Educational	Educational materials for high-school	47,805
	Educational materials for primary school	23,192
Scientific Art.	Scientific articles on specialized topics	471,979
	Wikipedia Art. (Ecology/Environment portal)	204,460
Narrative	Terence&Teacher-original versions	27,833
	Terence&Teacher-simplified versions	25,634

Dataset

Genre	Corpus	Tokens
Journalism	Repubblica	232,908
	DueParole	72,884
Educational	Educational materials for high-school	47,805
	Educational materials for primary school	23,192
Scientific Art.	Scientific articles on specialized topics	471,979
	Wikipedia Art. (Ecology/Environment portal)	204,460
Narrative	Terence&Teacher-original versions	27,833
	Terence&Teacher-simplified versions	25,634

Results

Distribution of “head–initial” and “head–final” syntactic pairs across the corpora



ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC

www.italianlp.it

Results

Linear distance between the dependent and the head in the canonical and non-canonical position

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

Results

Subordinate clause: distance, length and depth

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

Want to know more?

Visit our poster!