Studio sull'ordine dei costituenti nel confronto tra generi e complessità

Giulia Pieri•, Dominique Brunato◇, Felice Dell'Orletta◇
• Università di Pisa, Emm&mmE Informatica
giulia.pieri@mminformatica.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

Italiano. In questo articolo presentiamo uno studio sull'ordine dei costituenti in italiano basato su corpora annotati in maniera automatica fino all'analisi sintattica a dipendenze. L'indagine comparativa ha permesso di valutare l'influenza sia del genere testuale sia della complessità linguistica nella distribuzione dei fenomeni di marcatezza sintattica.

English. *In this paper we present a study on the order of constituents in Italian based on automatically dependency–parsed corpora. The comparative investigation has allowed to evaluate the influence of the textual genre and the linguistic complexity on the distribution of phenonemena of syntactic markedeness.*

1 Introduzione

Sebbene non esista una metrica universalmente valida con la quale poter classificare le lingue secondo una scala di complessità (McWorther, 2001), esistono alcuni indicatori che, a diversi livelli linguistici, possono essere assunti come indici di complessità 'universalmente' validi (Fiorentino, 2009). Sul piano sintattico, uno di essi è rappresentato dall'ordine dei costituenti, per cui le lingue che ammettono un ordine libero sono considerate più complesse di quelle a ordine fisso. Nella letteratura linguistica e psicolinguistica la flessibilità dell'ordine viene ricondotta, a sua volta, a fattori diversi che tengono in considerazione, da un lato, i principi semantici e pragmatici determinati dalla struttura dell'informazione (Diessel, 2005), dall'altro i vincoli di *performances*, per cui le strutture non marcate sono quelle cognitivamente meno costose che permettono al parlante di elaborare l'informazione più velocemente (Hawkins, 1994; Gibson, 1998; Gibson, 2000). Esaminando in maniera comparativa due treebank del Latino e del Greco antico, lo studio di (Gulordava e Merlo, 2015) ha dimostrato come la flessibilità dell'ordine sintattico, misurata come la distanza tra l'effettiva lunghezza delle dipendenze di una frase e la sua lunghezza ottimale (Gildea e Temperley, 2010), sia un elemento di complessità che si può desumere tanto dalla minor precisione del parsing automatico nell'analisi di queste lingue, quanto dalla tendenza che si riscontra nel tempo verso modelli di ordine fisso dei costituenti.

Questo articolo propone uno studio quantitativo per l'italiano, lingua di tipo VO (o a testa iniziale) e relativamente poco flessibile, volto a indagare se, e in che misura, la disposizione naturale o *non marcata* dei costituenti nella frase sia influenzata dal genere testuale e dalla complessità della lingua usata nel testo. A questo scopo sono stati comparati due generi linguistici, narrativo e giornalistico, a loro volta distinti in due varietà linguistiche differenti per grado di complessità, dove tale grado è definito in relazione al lettore di riferimento. A differenza delle analisi tradizionali di tipo *corpus–based* sull'ordine dei costituenti in italiano, tutti i dati qui discussi sono ricavati da corpora annotati in maniera automatica fino al livello di analisi sintattica a dipendenze. Anche se la ricostruzione della struttura sintattica da parte di un parser statistico è soggetta inevitabilmente ad alcuni errori (Montemagni, 2013), che aumentano per i testi di un dominio distante da quello del *training* (Gildea, 2001), la varietà dei fenomeni che si possono monitorare con affidabilità a partire da un'analisi linguistica automatica è molto ampia e complessa. La prospettiva linguistico–computazionale apre dunque prospettive di ricerca promettenti per la costruzione e la validazione su larga scala di modelli teorici sul funzionamento dei sistemi linguistici sia in chiave tipologica sia rispetto ai tradizionali assi di variazione linguistica.

In quanto segue, verranno prima presentati i corpora utilizzati in questo studio e successivamente la metodologia di monitoraggio sui cui si è basata l'estrazione delle caratteristiche linguistiche oggetto di indagine (Paragrafo 2.1); nel Paragrafo 3 discuteremo i principali risultati ottenuti e infine trarremmo alcune conclusioni di questa ricerca.

2 I corpora

I corpora esaminati appartengono a due diversi generi testuali, narrativo e giornalistico. Per ciascun genere sono state selezionate due collezioni di testi rappresentative di due varietà di lingua che si possono collocare a due poli opposti per complessità linguistica, dove il grado di complessità è definito in base al destinatario previsto; ogni macroraccolta, dunque, contiene una collezione di testi "complessi" e una di testi "semplici".

I due corpora narrativi, *Terence* e *Teacher*, rappresentano la prima risorsa italiana per lo studio della semplificazione automatica e semi-automatica dei testi (Brunato et al., 2015). Entrambi sono costituiti da testi nella versione originale e nella rispettiva versione semplificata, allineate per ciascun corpus a livello di frase. Le versioni semplificate derivano da due differenti strategie di semplificazione manuale: la strategia "strutturale", che implica una semplificazione cumulativa (ovvero su diversi livelli linguistici) prodotta da esperti nel caso di *Terence*, e la strategia "intuitiva", che si avvale invece dell'intuizione e dell'esperienza dell'insegnante nel caso di *Teacher*. In particolare, *Terence* si compone di 32 racconti brevi per l'infanzia e delle rispettive versioni semplificate rivolte a bambini dai 7 agli 11 anni con deficit uditivi o con difficoltà nella comprensione dei testi [1]. *Teacher* è un corpus formato da 24 coppie di testi originali e semplificati raccolti da siti web educativi specializzati che forniscono risorse gratuite per gli insegnanti; in questo caso, il target della semplificazione sono principalmente studenti di lingua italiana L2.

Per il genere giornalistico, invece, il materiale analizzato è costituito da due corpora che raccolgono rispettivamente testi esemplificativi di una varietà complessa, *Repubblica*, e di una varietà semplice, *Due Parole*. Il primo (Rep) consiste in un ampio corpus di testi giornalistici (pari a 232.908 tokens) che include tutti gli articoli pubblicati dal 2000 al 2005 sul quotidiano *La Repubblica*, che si rivolge ad una platea di lettori con un profilo culturale medio-alto. Il secondo (2Par) è un corpus di 73.314 tokens che trae il nome dall'omonimo quotidiano *Due Parole*, un mensile di facile lettura curato da linguisti esperti in semplificazione dei testi che hanno utilizzato un linguaggio controllato per un pubblico adulto con un basso livello di alfabetizzazione o con lievi disabilità intellettuali (Piemontese, 1996). Il corpus qui analizzato comprende tutti gli articoli scritti tra il 2001 e il 2006. È importante sottolineare che, a differenza dei corpora di narrativa, il corpus giornalistico non è parallelo, in quanto i relativi testi "semplici" (quelli di *Due Parole*) non sono il risultato di un processo di semplificazione dei testi originali di *Repubblica*.

2.1 Analisi linguistica dei corpora

Come passo preliminare allo studio dei fenomeni di ordinamento sintattico riportati in Sezione 3, i corpora sono stati arricchiti automaticamente con annotazione morfo–sintattica e sintattica utilizzando la catena di analisi linguistica *LinguA*[2], che integra il Part–of–Speech tagger descritto in (Dell'Orletta, 2009) e il parser a dipendenze DeSR (Attardi et al., 2009). L'annotazione linguistica multi–livello ha permesso di analizzare gli stessi tramite MONITOR–IT: questo strumento, adottando la metodologia di monitoraggio descritta in Montemagni (2013), consente di ricavare la distribuzione di un'ampia gamma di caratteristiche lessicali, morfo–sintattiche e sintattiche rintracciate automaticamente in un corpus a partire dall'output dei diversi livelli di annotazione linguistica.

3 Analisi dei dati

Per gli scopi di questa indagine sono di interesse caratteristiche di ordine sintattico che fanno riferimento alla posizione lineare di un elemento rispetto alla "testa" da cui è retto in una rappresentazione sintattica a dipendenze. Gli elementi considerati sono stati: il soggetto, l'oggetto, l'avverbio, l'aggettivo e la clausola subordinata, di cui sono state calcolate: i) le occorrenze nella posizione "canonica" rispetto alla matrice prevalente SVO dell'italiano (preposta o posposta alla testa a seconda dell'elemento indagato) e nella posizione opposta, dunque "marcata" sintatticamente e/o

[1] Questo corpus deriva dall'omonimo progetto dell'Unione Europea (Terence Consortium, 2012).

[2] http://linguistic-annotation-tool.italianlp.it/

Corpus	Oggetto				Soggetto				Aggettivo				Avverbio			
	Pre-V		Post-V		Pre-V		Post-V		Pre-N		Post-N		Pre-V		Post-V	
	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD
TT orig	9.18	1.93	90.82	2.52	85.38	2.56	14.62	2.88	53.91	1.11	46.09	1.2	55.49	2.4	44.51	1.61
Rep	8.37	2.43	91.63	2.72	80.14	3.87	19.86	3.45	41.87	1.19	58.13	1.32	56.11	2.66	43.89	1.47
TT sempl	7.87	1.93	92.13	2.43	84.28	2.23	15.72	2.63	56.53	1.12	43.47	1.16	56.24	2.19	43.76	1.47
2Par	3.47	1.6	96.53	2.56	89.11	3.07	10.89	3.5	24.97	1.09	75.03	1.12	56.69	3.84	43.31	1.4

Tabella 1: Ordine relativo dei costituenti (%) e distanza media (AvD) rispetto alla testa verbale (V) o nominale (N).

pragmaticamente; ii) la distanza (in numero di tokens) del dipendente dalla testa sintattica in entrambe le posizioni. Per ognuno di questi dati, il confronto tra i corpora è avvenuto su due livelli: la variazione di genere e il grado di complessità. Infatti, scopo dello studio è stato verificare quali sono gli ordini degli elementi che vengono condizionati dal genere testuale e quali dipendono dal grado di complessità: l'ipotesi di partenza era che fosse possibile ritrovare una somiglianza dell'ordine degli elementi in relazione al genere, ma soprattutto verificare che, indipendentemente dal genere, i testi semplici sono più fedeli a seguire l'ordine canonico degli elementi, mentre i testi complessi presentano una più alta percentuale di casi di ordine marcato.

La Tabella 1 mostra i risultati del monitoraggio relativi all'oggetto, al soggetto, all'aggettivo e all'avverbio[3]. Partiamo dall'analisi degli elementi che, nel confronto complessivo tra corpora, dimostrano una tendenza più netta a ricorrere nella posizione canonica: l'oggetto e il soggetto. Nel caso dell'oggetto, si osserva che i testi giornalistici si attengono maggiormente all'ordine canonico, mentre nei testi narrativi aumentano lievemente le occorrenze dell'oggetto in posizione preverbale. L'ordine marcato con anteposizione dell'oggetto alla testa verbale è inoltre influenzato dal grado di complessità della lingua: in ciascun genere infatti, quest'ordine ricorre in percentuale minore nei testi semplici e tale differenza è evidente soprattutto in *2Par* che registra poco più del 3% di oggetti in posizione preverbale. Anche rispetto alla posizione del soggetto, è possibile notare un'influenza sia del genere sia della complessità. In questo caso, però, sono i testi narrativi originali a rispettare maggiormente l'ordine canonico soggetto–verbo (85,38%) rispetto a quelli di *Repubblica* (80,14%). La variazione rispetto al grado di complessità produce invece risultati coerenti alle aspettative solo per la prosa giornalistica, dove lo scarto tra *Rep* e *2Par* è quasi di 10 punti percentuali in favore dell'ordine canonico (*2Par*: 89,11%). Al contrario, la semplificazione dei testi narrativi ha prodotto un aumento, seppure minimo, di soggetti postverbali (TT orig: 14,62%; TT semp: 15,72%). Pur considerando che i testi narrativi originali sono comunque più semplici di quelli di *Repubblica*, proprio perché rivolti a bambini, questo dato potrebbe segnalare che forme di marcatezza sintattica sono talvolta preferite come esito della semplificazione, perché permettono di ottenere un testo narrativo più coeso, mantenendone la progressione tematica. Interessanti sono anche i dati sulla distanza lineare tra soggetto e verbo che, in entrambi i generi della varietà semplice, aumenta quando il soggetto è in posizione postverbale. Si può ipotizzare, tuttavia, che la presenza dei tratti di accordo sul verbo in una lingua come l'italiano renda meno difficoltosa la ricostruzione della dipendenza soggetto–verbo, anche quando il soggetto è in posizione marcata.

A differenza del soggetto e dell'oggetto, l'aggettivo in italiano ha una posizione meno rigida nel sintagma nominale. Infatti, anche se la posizione non marcata è generalmente postnominale, essa varia in base alla funzione semantica che l'aggettivo svolge rispetto al nome (Cinque, 2010). Questa flessibilità trova conferme nell'analisi empirica, tuttavia con differenze rispetto al genere: i testi giornalistici, infatti, privilegiano l'ordine tendenzialmente non marcato mentre quelli narrativi mostrano la tendenza opposta. Anche in questo caso, sul piano della variazione testi complessi/testi semplici, l'effetto è marcato solo per il genere giornalistico (*Rep*: 58,13; *2Par*: 75,03).

Considerazioni analoghe possono essere avanzate per l'avverbio, la cui posizione in italiano, pur essendo tendenzialmente postverbale, gode di ampia flessibilità in relazione alla classe semantica di ap-

[3]Per rendere possibile il confronto tra gradi di complessità, i corpora Terence e Teacher sono stati uniti così da ottenere due corpora, l'uno composto di tutti i testi narrativi originali (TT orig), pari a 26.311 tokens, e l'altro di tutti i relativi testi semplificati (TT semp), pari a 24.083 tokens.

partenenza (Bonvino et al., 2008). In tutti e quattro i corpora è preferita la posizione preposta al verbo, che è anche quella a generare link sintattici mediamente più lunghi (si veda il dato riportato nella terzultima colonna). Si tratta di un dato significativo, soprattutto se si considera che il valore medio più elevato è riportato proprio dai testi di *2Par* (3.84 tokens). Come per il caso del soggetto, anche questo dato suggerisce la necessità di raffinare una nota misura di complessità sintattica quale la distanza dei link sintattici, tenendo in considerazione proprietà semantiche e morfologiche degli elementi coinvolti nella relazione di dipendenza.

Infine, abbiamo condotto uno studio più dettagliato sulla subordinazione (Tabella 2). Anche in questo caso sono state estratte sia le distribuzioni percentuali della subordinata in posizione preposta e posposta alla reggente sia la distanza (in numero di tokens) che separa la part–of–speech che introduce la subordinata [4] dal verbo della reggente. Inoltre, questo dato è stato ulteriormente raffinato andando a calcolare la lunghezza totale (in tokens) dell'intera clausola subordinata e la sua profondità media, quest'ultima computata come numero di relazioni di dipendenza che intercorrono tra la radice del sotto–albero della subordinata e una parola senza dipendenti (foglia).

Corpus	**Subordinata**			
	Pre-Principale			
	%	AvD	Length	Depth
TT orig	10.12	9.71	8.93	3.86
Rep	15.37	11.51	9.49	4.16
TT sempl	11.03	8.0	7.19	3.63
2Par	15.71	10.26	7.43	3.72
	Post-Principale			
TT orig	89.88	3.27	8.62	4.19
Rep	84.63	3.44	12.07	5.28
TT sempl	88.97	2.94	7.91	4.12
2Par	84.29	3.0	8.39	4.36

Tabella 2: Ordine della clausola subordinata rispetto alla principale. Per ciascuna posizione, vengono riportate la distribuzione percentuale (%), la distanza media dalla principale (AvD), la lunghezza media (Length) e la profondità media (Depth) dell'intera subordinata.

I risultati indicano una netta preferenza per la posizione posposta rispetto alla principale. Il dato è coerente con le previsioni dei modelli di *processing* secondo cui questo ordinamento comporta un impegno cognitivo minore da parte del parlante e dell'ascoltatore perché consente di minimizzare i domini di riconoscimento delle relazioni sintattiche (Hawkins, 1994). Anche se meno frequenti, i casi di anteposizione della subordinata si verificano maggiormente nel genere giornalistico, addirittura nella varietà semplice (*2Par*: 15.71% *Rep*: 15.37%). Questi dati sono riconducibili alle teorie che chiamano in causa l'interazione tra sintassi e fattori pragmatici e semantici, per cui il genere giornalistico sarebbe più propenso ad anteporre la subordinata alla principale poiché costituisce lo sfondo tematico dell'evento principale e conferisce la funzione di collegamento tematico e introduzione per l'informazione nuova (Diessel, 2005). Come prevedibile, l'anteposizione della subordinata determina dipendenze sintattiche mediamente più lunghe; la difficoltà di processing che ne deriva è compensata dall'uso di subordinate più semplici, non solo in termini di lunghezza totale ma soprattutto strutturalmente: in tutti i corpora, infatti, le "catene" subordinanti hanno una profondità media minore quando la subordinata precede la principale.

4 Conclusione

Questo articolo ha proposto uno studio comparativo su un particolare fenomeno relativo alla complessità sintattica, ovvero l'ordine dei costituenti in italiano. Il confronto è stato condotto su due livelli: la variazione di genere e il grado di complessità.

Per quanto riguarda il primo, è stato possibile constatare che i testi giornalistici sono quelli che maggiormente si attengono all'ordine canonico degli elementi, mentre i testi narrativi hanno riportato una frequenza superiore di ordini marcati. Dal punto di vista della complessità, è chiara la tendenza in entrambi i generi a utilizzare l'ordine canonico come esito della semplificazione, sia a seguito di un processo di semplificazione di un testo originale, sia quando il testo nativamente è concepito come testo semplice.

Indipendentemente dal genere, il fenomeno che è risultato più legato alla complessità riguarda l'uso delle subordinate. In entrambi i generi prevalgono nettamente subordinate posposte alla principale in quanto più facili da processare e quando questa posizione non è rispettata si registra una tendenza alla semplificazione della subordinata stessa sia in termine di numero di parole, ma soprattutto strutturalmente, in termini di profondità del sottoalbero sintattico.

[4] Sono state considerate sia le subordinate esplicite, introdotte da una congiunzione subordinante, sia quelle implicite, introdotte da un verbo di modo infinito o da una preposizione.

Va infine ricordato che tutte le osservazioni riportate in questo studio sono basate su testi linguisticamente annotati in maniera automatica, dunque soggetti a errore. Nonostante ciò, ci aspettiamo che almeno limitatamente all'analisi di testi dello stesso dominio e varietà di lingua, le distribuzioni degli errori siano simili, permettendo dunque un confronto interno rispetto ai parametri linguistici indagati. L'affidabilità dei dati discussi è inoltre corroborata dal fatto che sono stati considerati testi standard, linguisticamente vicini a quelli sui quali gli strumenti di annotazione automatica sono tipicamente addestrati. D'altra parte, proprio perché la distribuzione degli errori potrebbe variare al variare del dominio dei testi, tra gli sviluppi di questo lavoro intendiamo condurre delle analisi a campione per verificare l'impatto dell'errore sui confronti ottenuti rispetto alle diverse strutture esaminate.

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, Italia, Dicembre 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*.

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, Italia, Dicembre 2009.

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

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

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, 95–126.

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

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

Kristina Gulordava, 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 2015, pp. 121–130.

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

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

Simonetta Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, (1), 145-172.

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

Terence Consortium. 2012. Story simplification: User guide. Restricted Distribution.