Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessità

Erica Tusa
Università di Pisa
ericatusa@hotmail.it

Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{nome.cognome}@ilc.cnr.it

Abstract

English. In this work, we will investigate whether and to what extent algorithms typically used to assess the reliability of the output of syntactic parsers can be used to study the correlation between processing complexity and the linguistic notion of markedness. Although still preliminary, achieved results show the key role of features such as dependency direction and length in defining the markedness degrees of a given syntactic construction.

Italiano. *In questo lavoro indagheremo se e come algoritmi tipicamente utilizzati per valutare l'affidabilità dell'analisi prodotta da strumenti di annotazione sintattica automatica possono essere utilizzati per studiare la correlazione tra complessità computazionale e la nozione linguistica di marcatezza. I risultati raggiunti, sebbene ancora preliminari, mostrano il ruolo chiave di fattori quali l'orientamento della relazione e la lunghezza della dipendenza nel definire le varie "sfumature" di marcatezza di una stessa relazione.*

1 Introduzione

Fin dagli anni '80, è andata affermandosi la convinzione che metodi e tecniche sviluppate nell'ambito della linguistica computazionale potessero contribuire a far avanzare la ricerca fornendo nuova evidenza per lo studio di nozioni chiave della linguistica teorica. "Computational linguistics provides important potential tools for the testing of theoretical linguistic constructs and of their power to predict actual language use": così si apre il contributo di Kučera (1982), che fondandosi sulla correlazione tra marcatezza e frequenza propone i risultati di uno studio computazionale della nozione di marcatezza a livello lessicale e grammaticale condotto sul Brown corpus. Se da un lato l'evidenza raccolta conferma la correlazione attesa tra frequenza e marcatezza, dall'altro vengono registrati casi interessanti in cui "the statistical evidence from the Brown Corpus offers both greater problems and greater insight. [...] The frequency data is [...] the reverse of what one might have assumed under the markedness analysis".

Oggi, a più di 40 anni dallo studio pioniestico di Kučera (1982) che si fondava su un corpus annotato morfo–sintatticamente di circa un milione di parole, è possibile estrarre da corpora di ben maggiori dimensioni informazione linguistica accurata e variegata. L'affidabilità crescente degli strumenti di Trattamento Automatico del Linguaggio sta rendendo infatti possibile l'acquisizione di evidenza quantitativa e computazionale che spazia attraverso diversi livelli di descrizione linguistica, incluso quello sintattico, che può contribuire in modo significativo allo studio di questioni aperte della letteratura linguistica: Merlo (2016) rappresenta un importante esempio di questo rinnovato filone di studi.

All'interno del quadro delineato sopra, l'obiettivo del presente contributo è verificare se e in che misura algoritmi per l'identificazione dell'affidabilità e plausibilità dell'annotazione sintattica possano contribuire allo studio di un fenomeno linguistico quale la marcatezza. Con l'uso sempre più diffuso dell'annotazione sintattica a dipendenze come punto di partenza per una vasta gamma di applicazioni finalizzate all'estrazione di informazione da vaste collezioni documentali, tali algoritmi nascono dalla necessità di identificare al-

l'interno delle annotazioni prodotte in modo automatico quelle corrette o, più semplicemente, caratterizzate da un maggior grado di affidabilità e plausibilità. Questo tipo di valutazione può avvenire in relazione sia all'intero albero sintattico assegnato alla singola frase (cfr. ad esempio Dell'Orletta *et al.* (2011) e Reichart and Rappoport (2009b)), sia alla singola relazione di dipendenza (si vedano, tra gli altri, Dell'Orletta *et al.* (2013) e Che *et al.* (2014)). Se da un lato l'identificazione di alberi sintattici corretti rappresenta un ingrediente fondamentale all'interno di algoritmi di Active Learning (Settles, 2012) o di apprendimento automatico semi–supervisionato e non supervisionato (Goldwasser *et al.*, 2011), l'identificazione dell'affidabilità di singole relazioni di dipendenza e/o sotto–alberi sintattici diventa fondamentale, ad esempio, per fornire evidenza utile a migliorare le prestazioni di un sistema di analisi sintattica automatica (vanNoord, 2007; Chen *et al.*, 2009), oppure per l'estrazione di nuclei di informazione affidabili.

Nel presente studio, ci focalizzeremo sul secondo tipo di algoritmi, ovvero quelli che operano a livello della singola relazione di dipendenza, per verificarne le potenzialità nello studio della nozione di marcatezza linguistica. Per quanto questi algoritmi operino tipicamente su corpora annotati in modo automatico (Dickinson, 2010), sono attestati usi anche su corpora con annotazione validata manualmente (qualificata come "gold"): in questo caso, il fine consiste nell'identificazione di errori e incoerenze di annotazione (Dickinson, 2015). Il risultato di questi algoritmi varia da una classificazione binaria della dipendenza (corretta $vs.$ errata) come in Che *et al.* (2014), a un ordinamento delle relazioni secondo l'affidabilità e la plausibilità dell'analisi, come proposto da Dell'Orletta *et al.* (2013). Al di là di differenze a livello dell'algoritmo utilizzato e del tipo di risultato, in tutti i casi viene fatto uso di un esteso inventario di caratteristiche linguistiche selezionate come indicatori di complessità.

2 L'ipotesi di ricerca

Combinando la prospettiva linguistica e quella linguistico–computazionale, l'ipotesi che intendiamo esplorare è se il punteggio assegnato da algoritmi per la misura della plausibilità dell'annotazione possa essere utilizzato per ricostruire il passaggio graduale da costruzioni non marcate a costruzioni caratterizzate da gradi crescenti di marcatezza. L'assunto di base sottostante a tale ipotesi si fonda sulla correlazione, ampiamente adottata nella letteratura linguistica, tra marcatezza e complessità: se da un lato costruzioni non marcate saranno caratterizzate da un maggior livello di plausibilità di annotazione (dunque da un minore livello di complessità), dall'altro costruzioni caratterizzate da gradi crescenti di marcatezza saranno associate a punteggi di minore di plausibilità (equivalente a una maggiore complessità).

All'interno della letteratura linguistica, la "marcatezza" rappresenta una nozione ampiamente dibattuta e altamente polisemica. Secondo quanto affermato da Haspelmath (2006), a partire dalle prime accezioni delineate negli anni '30 (Jakobson, 1932) essa può essere ricondotta a dodici significati diversi, organizzati in quattro classi: "markedness as complexity, as difficulty, as abnormality, or as a multidimensional operation". Tra queste, il presente contributo intende focalizzarsi sulla definizione di "markedness as abnormality" e, in particolare, sull'idea che quando consideriamo "marcato" un determinato evento linguistico lo stiamo considerando "abnormal", ovvero deviante rispetto a strutture linguistiche riconosciute come basiche all'interno della "norma linguistica". In questa ottica, la marcatezza come devianza rispetto alla norma è strettamente connessa sia con la frequenza d'uso (cfr. "markedness as rarity in texts"), sia rispetto alla distribuzione di un evento linguistico all'interno di una varietà più o meno ampia di contesti linguistici (cfr. "markedness as restricted distribution").

In quanto segue, analizzeremo i risultati di un algoritmo per la valutazione della plausibilità di singole relazioni alla luce delle accezioni di marcatezza selezionate.

3 Metodologia di analisi e corpora

L'algoritmo che abbiamo utilizzato per la misura della plausibilità dell'annotazione sulla base della quale produrre l'ordinamento delle relazioni di dipendenza è costituito da LISCA (Dell'Orletta *et al.*, 2013). Tale algoritmo assegna a ogni relazione – definita come una tripla (d, h, t) dove d è il dipendente, h è la testa, e t è il tipo di dipendenza che connette d a h – un valore di *plausibilità*. LISCA opera in due fasi: 1) colleziona statistiche relative a un insieme di caratteristiche linguistica-

mente motivate estratte da un ampio corpus di alberi a dipendenze ottenuti attraverso un processo di annotazione sintattica automatica; 2) combina queste statistiche all'interno di una funzione descritta in Dell'Orletta *et al.* (2013) per ottenere il punteggio da associare all'arco sintattico in corso di valutazione. La combinazione viene calcolata come il prodotto dei pesi associati a ciascuna caratteristica identificata.

La Figura 1 descrive graficamente le caratteristiche prese in esame da LISCA per la misura della plausibilità dell'arco sintattico (d, h, t). Ai fini del presente studio, LISCA è stato utilizzato nella sua variante delessicalizzata per poter fare astrazione da variazioni di natura lessicale. In particolare, sono stati presi in considerazione due diversi tipi di caratteristiche, entrambe associate nella letteratura linguistica alla nozione di complessità sintattica:

- tratti *locali*, corrispondenti alle peculiarità dell'arco sintattico considerato, come ad esempio la distanza in termini di tokens all'interno della frase tra d e h, oppure la forza associativa che unisce le categorie grammaticali coinvolte (POS_d e POS_h), o la POS della testa di h e il tipo di relazione sintattica che li lega;

- tratti *globali*, volti a localizzare l'arco considerato all'interno della struttura sintattica della frase, ad esempio la distanza di d rispetto alla radice dell'albero, oppure rispetto alla foglia più vicina o a quella più lontana, oppure il numero di nodi "fratelli" e "figli" di d ricorrenti rispettivamente alla sua destra–sinistra nell'ordine lineare della frase.

In questo studio, per estrarre le statistiche rispetto alle caratteristiche linguistiche prese in esame, LISCA è stato applicato a un corpus di 1.104.237 frasi (22.830.739 tokens) estratte da articoli del quotidiano *La Repubblica*, parte del CLIC-ILC Corpus (Marinelli et al., 2003). Il corpus è stato annotato a livello morfosintattico con l'ILC–POS–Tagger (Dell'Orletta *et al.*, 2009) e a livello sintattico a dipendenze con DeSR (Attardi *et al.*, 2009). Gli strumenti di annotazione sono stati addestrati sulla *Italian Universal Dependecies Treebank*, in breve IUDT (Bosco *et al.*, 2013). Lo schema di annotazione utilizzato è quello delle "Universal Dependencies", concepito per massimizzare il parallelismo delle annotazioni in lingue diverse e che per questo motivo privilegia relazioni

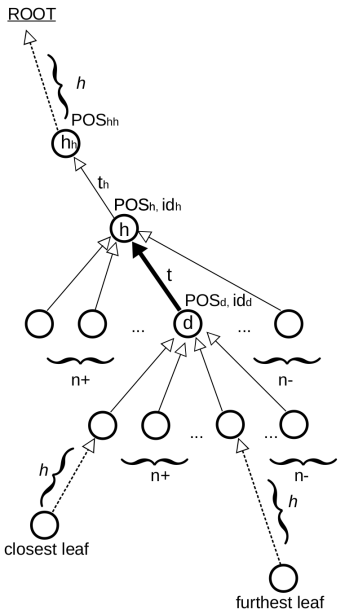

Figura 1: Caratteristiche utilizzate LISCA per il calcolo della plausibilità dell'arco (d, h, t).

di dipendenza tra parole lessicali trattando le parole grammaticali come dipendenti di parole semanticamente piene (Nivre, 2015). IUDT costituisce anche il corpus di indagine di questo lavoro: attraverso LISCA a ogni arco sintattico del corpus è stato assegnato un punteggio.

4 Analisi dei dati

I punteggi assegnati da LISCA alle relazioni di dipendenza della IUDT sono stati utilizzati per ordinare le relazioni in ordine descrescente di plausibilità. La lista ordinata così ottenuta è stata suddivisa in 10 fasce di 24,644 relazioni ciascuna (corrispondente al 10% del totale). Partendo dall'analisi della variazione della distribuzione dei tipi di relazioni di dipendenza attraverso le fasce, ci siamo poi focalizzati su singole relazioni, con l'intento di ricostruire il passaggio da non marcato (o prototipico) a marcato, e di identificare i fattori che contribuiscono a determinare il grado di marcatezza di una costruzione, definita in questo studio come una relazione di dipendenza all'interno del contesto sintattico di occorrenza. Nella consapevolezza che le relazioni sono distribuite all'interno delle fasce in virtù della combinazione di tutte le caratteristiche locali e globali prese in considerazione, ci siamo focalizzati su due parametri ampiamente indagati nella letteratura linguistica con l'intento di verificare se e in che misura l'ordinamento di

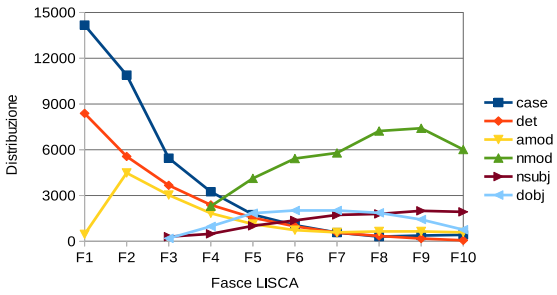

Figura 2: Distribuzione di una selezione di dipendenze.

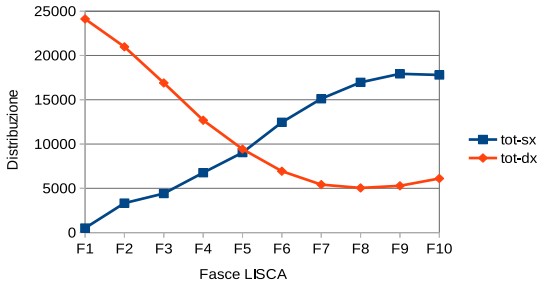

Figura 3: Orientamento generale delle dipendenze.

LISCA rifletta note gerarchie di marcatezza. In particolare, è stato analizzato il ruolo i) dell'orientamento della dipendenza, definito dalla direzione verso destra o sinistra dell'arco sintattico che lega d a h rispetto all'ordine lineare delle parole nella frase, e ii) della lunghezza della relazione di dipendenza, calcolata come la distanza in parole tra d e h.

4.1 Distribuzione delle dipendenze

Partiamo dall'analisi della distribuzione delle dipendenze nelle 10 fasce. Nelle prime fasce si osserva un insieme ristretto di tipi di dipendenze, la cui frequenza diminuisce proporzionalmente al decrescere dei punteggi assegnati da LISCA. Man mano che si prosegue verso le fasce intermedie, i tipi di dipendenza all'interno di ciascuna fascia si fanno più numerosi e variabili. Gli archi con i punteggi più alti, che si collocano nelle prime fasce, mettono in relazione parole grammaticali e parole lessicali, come *det*(erminer), *case* e *mark*(er), che collegano, rispettivamente, articoli, preposizioni, congiunzioni subordinanti o avverbi alla relativa testa. Proseguendo oltre le prime due fasce, troviamo relazioni come *advmod* (adverbial modifier), *nummod* (numerical modifier), *cop*(ula) e *aux*(iliary) che collegano, rispettivamente, avverbi, numerali, copule e ausiliari (verbi modali compresi) alla loro testa. A partire dalla quinta fascia si osserva un'incidenza sempre maggiore di relazioni che collegano parole lessicali come sostantivi e verbi alla loro testa, ad esempio *nsubj* (nominal subject), *dobj* (direct object), *ccomp* (clausal complement), *xcomp* (clausal complement with controlled subject) e *root*. Un caso a parte è rappresentato dalle relazioni *amod* e *nmod* che collegano modificatori aggettivali o nominali con la relativa testa: esse si distribuiscono in modo simile in tutte le fasce.

La Figura 2 riporta l'andamento della distribuzione di sei relazioni di dipendenza selezionate come segue: due relazioni concentrate principalmente nelle prime fasce (*det*, *case*), due relazioni caratterizzate da una distribuzione più diffusa (*amod*, *nmod*) e due relazioni maggiormente ricorrenti nelle ultime fasce (*nsubj*, *dobj*). La distribuzione delle relazioni che vedono una parola grammaticale come dipendente all'interno delle prime fasce può essere ricondotta alle strutture generalmente fisse o poco variabili, che le rendono facilmente trattabili computazionalmente. D'altro lato, le parole lessicali tendono ad inserirsi in costruzioni più complesse, caratterizzate da una maggiore flessibilità a livello dell'ordine lineare all'interno della frase, e potenzialmente soggette a condizionamenti di tipo pragmatico che portano alla formazione di strutture sintattiche più complesse. La presenza diffusa della relazione *amod* attraverso le fasce rappresenta un caso diverso, non riconducibile alla libertà di movimento ma piuttosto alla direzione degli archi sintattici che li collegano alla loro testa: mentre la distanza media tra d e h in *amod* rimane tendenzialmente costante attraverso le fasce, la direzione della relazione varia significativamente. In quanto segue, ci concentreremo su due dei parametri che sembrano svolgere un ruolo chiave nella distribuzione delle relazioni attraverso le fasce.

4.2 Orientamento delle dipendenze

La Figura 3 riporta la distribuzione attraverso le fasce di tutte le relazioni di dipendenza, facendo distinzione tra dipendenze con testa a destra ($d > h$) e dipendenze con testa a sinistra ($h < d$). Si osserva che i due tipi di orientamento, nonostante ricorrano con frequenza molto simile (112.886 $d > h$ vs 104.301 $h < d$), sono descritti da andamenti opposti: nelle prime fasce si concentrano le

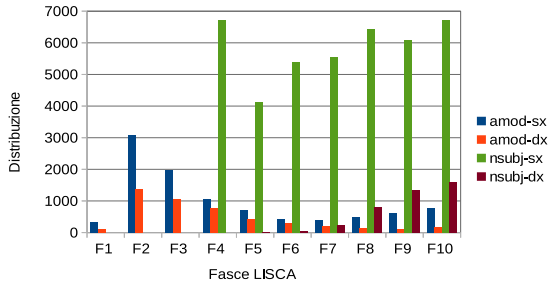

Figura 4: Orientamento di una selezione di dipendenze.

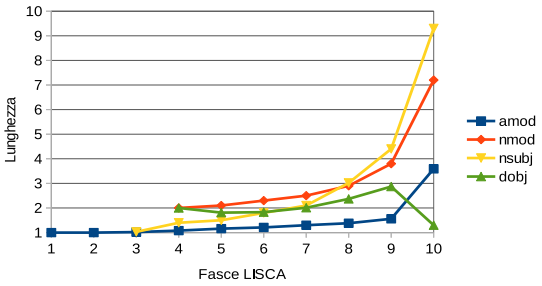

Figura 6: Andamento della lunghezza media di una selezione di dipendenze per fascia.

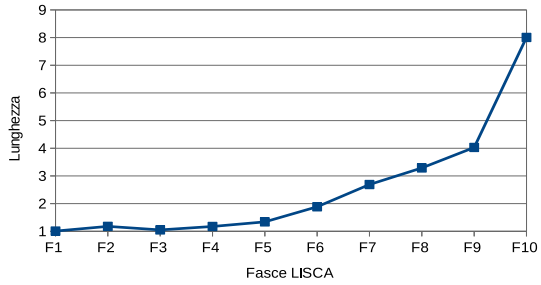

Figura 5: Andamento generale della lunghezza media delle dipendenze per fascia.

relazioni di tipo $d > h$, nelle ultime quelle $h < d$.

Nella Figura 4 è riportato l'andamento attraverso le fasce dell'orientamento di due relazioni, *amod* e *nsubj*. Nel caso di *amod* le teste dei modificatori aggettivali si trovano in netta maggioranza a sinistra, soprattutto nella seconda fascia: gli aggettivi postnominali, dunque in posizione non marcata, sono stati valutati come più plausibili e computazionalmente trattabili. Invece, *nsubj*, che collega il soggetto nominale alla testa verbale, nella maggior parte dei casi presenta la testa a destra: il soggetto preverbale corrisponde all'ordine non marcato in italiano. La sequenza verbo–soggetto è attestata, con andamento crescente, a partire dalla fascia 6. Le relazioni che sono state valutate con i punteggi più alti, ovvero *det* e *case*, mostrano la testa sempre a destra.

4.3 Lunghezza delle dipendenze

Nella Figura 5, per ciascuna fascia è riportata la media delle lunghezze delle relazioni di dipendenza: si osserva che il punteggio di LISCA decresce in maniera inversamente proporzionale al valore della lunghezza media. Nella Figura 6 sono riportate le medie all'interno di ogni fascia delle lunghezze di un gruppo selezionato di dipendenze: alcune delle relazioni che abbiamo visto concentrarsi nelle prime fasce, come *det* e *case*, ma anche relazioni dalla distribuzione più bilanciata come *amod* e *dobj* risultano essere quelle più brevi (la loro lunghezza in media non supera una distanza di 2 parole tra testa e dipendente); dipendenze come *nsubj* e *nmod*, che collegano unità dall'ordinamento più flessibile, cioè caratterizzate da una maggiore libertà di movimento rispetto alla loro testa, raggiungono in media, soprattutto verso le ultime fasce, le lunghezze maggiori. Considerando il ruolo ampiamente ascritto in letteratura, non solo linguistico–computazionale ma anche linguistica e psicolinguistica, alla lunghezza delle dipendenze come fattore di complessità linguistica, questi dati costituiscono una prova ulteriore che l'ordinamento prodotto da LISCA riflette la marcatezza della costruzione.

5 Conclusione

In questo studio abbiamo esplorato l'ipotesi che algoritmi sviluppati per valutare l'affidabilità e la plausibilità di annotazioni sintattiche a dipendenze possano fornire evidenza utile a una riflessione attorno al tema della complessità sintattica, e in particolare a ricostruire "sfumature" di marcatezza crescente in relazione alla stessa relazione di dipendenza. I primi risultati raggiunti sono incoraggianti: quanto osservato in relazione alla distribuzione delle dipendenze attraverso le fasce ci porta a ipotizzare una forte correlazione tra la complessità computazionale dell'analisi individuata da LISCA e la nozione di marcatezza sintattica. È stato indagato in particolare il ruolo di fattori quali l'orientamento della relazione e la lunghezza della dipendenza, con risultati che mostrano chiaramente che un algoritmo come LISCA può essere un valido strumento anche per analisi di tipo linguistico. Attraverso l'analisi della distribuzione delle

relazioni di dipendenza nelle fasce definite sulla base dell'ordinamento di LISCA è stato possibile non solo discriminare tra costruzioni marcate e non marcate (dato tipicamente recuperabile sulla base della frequenza), ma anche identificare – relazione per relazione – i fattori che hanno contribuito a renderla marcata. Se l'orientamento della relazione gioca un ruolo cruciale nel caso di *amod*, nel caso di *nsubj* è piuttosto la distanza tra la testa e il dipendente a determinare la marcatezza della costruzione. Ovviamente, questa metodologia di analisi dovrebbe essere estesa alla vasta tipologia di caratteristiche linguistiche considerate.

Gli sviluppi correnti di questo lavoro includono: l'estensione, al di là della lunghezza e l'orientamento della relazione, della tipologia di fattori linguistici esplorati, per arrivare anche allo studio dell'impatto di fattori lessicali; l'estensione della tipologia di costruzioni analizzate, che potrebbero anche includere combinazioni di dipendenze corrispondenti a sotto–alberi sintattici. Riteniamo che la metodologia dovrebbe anche essere applicata a treebank di lingue diverse, così come diversi generi testuali all'interno della stessa lingua.

References

Attardi G., Dell'Orletta F., Simi M., Turian J. 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.

Bosco C., Montemagni, S., Simi, M. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August 2013.

Che W., Guo J., Liu T. 2014. ReliAble dependency arc recognition. In *Expert Systems with Applications*. volume 41, number 4, pp. 17161722.

Chen W., Kazama J., Uchimoto K., Torisawa K. 2009. Improving Dependency Parsing with Subtrees from Auto-parsed Data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP '09). Association for Computational Linguistics. volume 2, pp. 570–579.

Dell'Orletta F. 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.

Dell'Orletta F. 2011. ULISSE: an Unsupervised Algorithm for Detecting Reliable Dependency Parses. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011*, pp. 115–124.

Dell'Orletta F., Venturi G., Montemagni S. 2013. Linguistically-driven Selection of Correct Arcs for Dependency Parsing. In *Computaciòn y Sistemas*. ISSN 1405-5546, vol. 17, No. 2, pp. 125-136.

Dickinson M. 2010. Detecting Errors in Automatically-Parsed Dependency Relations. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, Association for Computational Linguistics, pp. 729–738.

Dickinson M. 2015. Detection of annotation errors in corpora. In *Language and Linguistics Compass*. ISSN 1749-818X, vol. 9, No. 3, pp. 119-138.

Goldwasser D., Reichart R., Clarke J., Roth D. 2011. Confidence Driven Unsupervised Semantic Parsing. In Proceedings of the *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA* (ACL–2011). pp. 1486–1495.

Haspelmath M. 2006. Against markedness (and what to replace it with). In *Journal of Linguistics*. ISSN 1469-7742, vol. 42, No. 01, pp. 25–70.

Jakobson R. 1932. Zur Struktur des russischen Verbums. In *Charisteria Gvilelmo Mathesio*.

Kučera H. 1982. Markedness and Frequency: a Computational Analysis. In *Proceedings of COLING 82*. pp. 167-173.

Marinelli R., L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari, A. Zampolli. 2003. *The Italian PAROLE corpus: an overview*. In Zampolli A. et al. (eds.), Computational Linguistics in Pisa, Special Issue, XVI–XVII, Pisa-Roma, IEPI. Tomo I, pp. 401-421.

Merlo P. 2016. Quantitative computational syntax: some initial results. In *Italian Journal of Computational Linguistics*. vol. 2.

Nivre J. 2015. Towards a Universal Grammar for Natural Language Processing. In *Proceedings of the 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Part I.* pp. 3–16.

Reichart Roi and Ari Rappoport. 2009b. *Sample Selection for Statistical Parsers: Cognitively Driven Algorithms and Evaluation Measures*. In Proceedings of CoNLL 2009, pp. 3–11.

Settles B. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers.

van Noord G. 2007. In *Proceedings of the 10th International Conference on Parsing Technologies* (IWPT–2007). Association for Computational Linguistics, pp. 1–10.