UNIVERSITÀ DI PISA

Dipartimento di Filologia, Letteratura e Linguistica

Corso di Laurea triennale in
Informatica Umanistica

RELAZIONE

Verso una nozione computazionale di complessità sintattica: un'analisi linguistico-computazionale del passaggio da *prototipico* a *marcato*

Candidato: *Erica Tusa*

Relatore: *Dott. Felice Dell'Orletta*

Correlatore: *Dott.essa Simonetta Montemagni*

Anno accademico 2014-2015

Indice

Introduzione

Il presente lavoro si inserisce nell'ambito di ricerca del *parsing a dipendenze* e si presenta come uno studio innovativo di complessità sintattica, proponendo una prospettiva d'analisi volta ad una nuova definizione computazionale di tale nozione e alla valutazione della complessità delle strutture morfo-sintattiche e sintattiche del linguaggio italiano scritto. Nello specifico la prospettiva d'analisi verrà fatta coincidere con quella di un algoritmo di *evaluation* (valutazione), utilizzato per misurare il grado di plausibilità e di prototipicità delle strutture sintattiche. Prima di giungere alle analisi elaborate nel presente studio, osserveremo nel dettaglio le premesse teoriche necessarie alla loro comprensione.

Per lo svolgimento di questo progetto è stata adottata una particolare metodologia di *monitoraggio* linguistico resa possibile grazie al ricorso di tecnologie linguistico-computazionali. Diversi recenti studi hanno dimostrato infatti come l'applicazione delle nuove tecnologie abbia reso possibile diversi compiti di ricerca e di monitoraggio, volti alla ricostruzione del profilo linguistico di campioni di testo, in relazione ai vari livelli di descrizione linguistica (primariamente, lessico, morfo-sintassi e sintassi) e alle dimensioni di variazione sociolinguistica (diacronia, diamesia, diafasia, diastratia). In diversi scenari applicativi il supporto fornito da sistemi computazionali ha dimostrato un forte potenziale analitico e descrittivo, permettendo analisi sempre più accurate ed affidabili, che coprono aspetti della struttura linguistica fino ad ora poco esplorati in quanto difficilmente indagabili mediante un'analisi manuale del testo. Come illustrato in (Montemagni, 2013) le diverse potenzialità offerte dall'uso di tecnologie linguistico-computazionali sono state confermate in un recente filone di studi avviato a livello internazionale, all'interno del quale le analisi linguistiche realizzate attraverso l'applicazione di strumenti di trattamento automatico del linguaggio sono state variamente utilizzate per il monitoraggio di un ampio spettro di tratti strutturali riguardanti tutte le principali dimensioni di variazione linguistica.

Nel dominio della sintassi, in particolare, metodi di monitoraggio facenti ricorso a tecnologie linguistico-computazionali sono stati fino ad ora applicati, con finalità di *information-extraction*, in diversi studi su argomenti di interesse linguistico, come le varietà di lingua, i generi testuali, la specificità dei linguaggi settoriali, l'opposizione tra linguaggio parlato e scritto, le tendenze e trasformazioni della lingua in prospettiva sia sincronica sia diacronica.

La tipologia di parametri che possono essere monitorati a partire dai livelli di annotazione

linguistica morfo-sintattica e sintattica a dipendenze, è molto ampia e varia, oltre che in costante espansione, in parallelo ad un incremento di disponibilità di risorse e allo sviluppo ed affinamento delle tecnologie computazionali. Utilizzando l'analisi sintatica automatica, infatti, è possibile rilevare importanti informazioni sulla struttura linguistica, come ad esempio la tipologia degli elementi coinvolti nelle varie costruzioni sintattiche, i livelli di incassamento gerarchico, l'ordinamento dei costituenti ecc. Inoltre i parametri derivati dalle caratteristiche strutturali dell'albero sintattico, comea ad esempio la "misura" della sua profondità o la "misura" della lunghezza delle relazioni di dipendenza, calcolata come la distanza (in parole) tra la testa e il dipendente, rivestono un ruolo centrale nella valutazione della complessità di un testo. Lo studio qui presentato, che si interessa appunto di complessità luinguistica, ne è un esempio.

Nel panorama culturale moderno la nozione di complessità sintattica rappresenta uno dei temi di ricerca teorica e applicativa tra i più discussi, e questo lo dimostra il numero di discipline che la comprendono nel loro dominio di interessi: dalla matematica, alla teoria dell'informazione, alle scienze computazionali e infine sociali. Sebbene sia ancora altamente dibattuta la possibilità di individuare una metrica universalmente valida con la quale poter classificare le lingue secondo una scala di complessità, diversi sono i framework teorici in cui essa è stata definita ed utilizzata.

La complessità viene generalmente analizzata in due modi: come complessità nel sistema e come complessità per l'utente.

- Complessità nel sistema: la complessità viene fatta dipendere dalla comparazione di sistemi linguistici e strutture linguistiche sulla base di criteri interni alle lingue (ad esempio si considerano il numero di regole per produrre un certo *output*, il numero di eccezioni alle regole, il numero di unità previste in un certo livello linguistico, la mancanza di trasparenza nella relazione forma significato, ecc.). In questo caso, dunque, la complessità viene definita dalla descrizione del linguista, che stabilisce ciò che è più complesso.
- Complessità per l'utente: la complessità viene definita considerando le restrizioni cognitive che l'utente impone alla forma della lingua, e viene a dipendere dalle strategie cognitive messe in atto dall'utente nei compiti di produzione e ricezione del linguaggio. L'approccio infatti valuta il modo in cui il linguaggio viene *processato* e dunque 'più complesso' è ciò che richiede più tempo per essere compreso, più passaggi nel processamento, un carico maggiore per la memoria di lavoro e un compito cognitivo più costoso. "Complesso" equivale dunque a "più difficile da comprendere e produrre".

Sia la complessità interna al sistema sia la complessità per l'utente sono considerate nozioni relative, non assolute, e generalmente vengono associate alla intensificazione di un tratto, come, ad esempio, più materiale linguistico, più alternative di variazione per una stessa funzione, più regole

per generare un output, minore trasparenza nella relazione forma-funzione, minore naturalezza/ maggiore marcatezza, più sintassi che pragmatica, più tempo di comprensione, più passaggi nel processing, più memoria di lavoro ecc.

Nell'ambito della sintassi una possibile definizione è stata proposta dalla classificazione di Gaetano Berruto e Massimo Cerruti (1997) secondo cui la complessità sintattica è da considerarsi come una proprietà linguistica inerente alla natura del sistema linguistico e alla sua complessità strutturale. Secondo questa classificazione la definizione di complessità viene ricollegata ad alcuni aspetti particolarmente rilevanti della struttura sintattica (come un sintagma o una frase), quali:

- l'**ordine** lineare degli elementi di una frase, che permette di evitare le possibili ambiguità di significato;
- le **dipendenze** tra gli elementi di una frase che evidenziano i rapporti gerarchici che essi hanno. Questi ultimi costituiscono una seconda trama della sintassi, sovrapposta alla successione lineare e del tutto indipendente da essa.
- le **incassature** di alcuni elementi all'interno di altri per indicare particolari legami e i rispettivi livelli in cui si trovano le diverse parti della catena linguistica;
- le **parti del discorso**, che danno informazioni sulla sua strutturazione interna.

Un importante cardine attorno a cui ruota la questione della complessità sintattica, soprattuto per quanto riguarda l'ordine dei costituenti, è la nozione di *marcatezza*. Questa si riferisce a un'opposizione tra due elementi linguistici, per la quale essi sono uguali in tutto salvo una peculiarità, detta appunto *marca*, che è presente in uno di essi e manca nell'altro. Se si classificano o descrivono le lingue in base all'ordine dei costituenti delle diverse strutture, si ammette che questi hanno una disposizione naturale, o *non marcata*, e una serie più o meno ampia di costruzioni 'devianti' (dette *marcate*) rispetto a essa, che hanno l'effetto di aggiungere un tratto (la marca) alla frase non marcata.

In sintassi una frase si dice marcata sintatticamente quando i *costituenti* che la compongono non occupano le loro posizioni canoniche, ma sono dislocati al fine di focalizzare una particolare informazione: la marcatezza può essere interpretata in chiave sia puramente sintattica, secondo i parametri definiti da Berruto e Cerruti, sia in chiave pragmatica, considerando i particolari scopi comunicativi soggiacenti al messaggio linguistico.

Utilizzando come parametro le posizione nella frase dei costituenti soggetto (S), oggetto diretto (O) e verbo (V) (Greenberg,1962), nella lingua italiana, l'ordine degli elementi viene definito "non marcato" quando il posizionamento dei costituenti corrisponde all'ordine basico SVO, secondo il principio di *testa a sinistra* (o *testa iniziale*), come riassunto nella Tabella 1. Tuttavia, per esigenze comunicative di natura pragmatica, spesso si è soliti ricorrere a un cambiamento dell'ordine dei

costituenti, allo scopo di focalizzare l'attenzione sull'informazione che ci preme prima comunicare: in questo caso, l'ordine viene definito "marcato". E' importante sottolineare come l'ordine canonico dell'italiano, che si propone come efficace strumento di previsione sulla struttura dei sistemi linguistici, si riferisca ad un ideale linguistico: la lingua italiana, infatti, presenta molteplici costruzioni sintattiche devianti rispetto all'ordine canonico (le parole lessicali hanno una notevole libertà di movimento che permette diverse forme di focalizzazione, mentre è più ridotta la libertà di spostamento delle parole grammaticali).

Funzione grammaticale	Soggetto	Verbo/Predicato verbale	Oggetto
Ruolo semantico	Agente	Azione	Paziente
Struttura informativa	Tema	Rema	
Ruolo pragmatico	Dato	Nuovo	

Tabella 1. Ordine non marcato dei costituenti in italiano.

In sintassi, riguardo l'ordine dei costituenti in una frase, si delineano due principali linee di pensiero: mentre alcuni studiosi opinano che l'ordine degli elementi linguistici, per una questione di efficacia comunicatica maggiore (Diessel, 2005), sia determinato dalla struttura dell'informazione di tipo "dato/nuovo", secondo cui il soggetto corrisponde nella maggior parte dei casi all'*informazione data*, mentre nel resto della frase si trasmette l'*informazione nuova*; altri ritengono che l'ordine sia in qualche modo dipendente dalla capacità computativa di riconoscere ed elaborare più velocemente l'informazione (Hawkins, 1994).

Al di là delle diverse teorie interpretative è indubbio che tutti i tratti che contribuisco a denifire la complessità fino ad ora citati rappresentano pametri cruciali negli studi linguistici, da quelli di linguistica descrittiva, a quelli di sociolinsuistica, linguistica tipologica e comparata, e infine linguistica computazionale. La svolta rappresentata dall'ausilio di sistemi computazionali automatici consiste nel provare a rendere possibile da una parte, il rilevamento di questi stessi parametri e, dall'altra l'analisi, in isolamento o in funzione a un sistema, delle strutture sintattiche del linguaggio, consentendo tempi di esecuzione più rapidi, e ricerche agevolate oltre che più approfondite. Negli studi di tipologia riguardanti le variazioni tra le lingue, ad esempio, soprattutto consideranto lingue come l'italiano, senza un sistema di casi e in cui l'ordine dei costituenti è relativamente poco flessibile, poter rilevare automaticamente i ruoli sintattici e le strategie formali utilizzate per realizzare la garammatica facilita notevolmente le attività del linguista.

La prospettiva di questo studio si presenta come innovativa in quanto, prendendo spunto dalle metodologie d'analisi sopra citate, tratterà la nozione di complessità in termini computazionali: a partire da una metrica capace di misurare il grado di plausibilità delle strutture sintattiche riscontrate in un corpus di testi correttamente annotati, si potrà osservare a livello statistico quali particolari

aspetti della grammatica vengono riconosciuti da un algoritmo di calcolo come *plausibili* e *prototipici*, e quali invece sono più complessi e più marcati.

Per poter elaborare queste "misure" sono state utilizzate precise risorse e tecniche computazionali, selezionate accuratamente in riferimento al grado di accuratezza ed affidabilità attestato in letteratura, di fondamentale importanza al fine di ottenere risultati statisticamente validi e rilevanti.

Prima di osservare i risultati ottenuti (capitolo 4), nei prossimi capitoli verranno introdotte le premesse teoriche necessarie alla compresione dei dati (capitolo 1), successivamente verrà fornita una panoramica delle risorse e strumenti computazionali utilizzati (capitoli 2-3), descrivendo nello specifico le tre fasi preparative all'elaborazione dei dati.

In particolare nel prossimo capitolo verrà introdotto a grandi linee l'ambito entro cui si iscrive il presente progetto, ovverro quello del parsing a dipendenze.

1 Grammatica e Parsing a dipendenze

1.1 Introduzione

In informatica il *parsing* è il processo di analisi volto a rilevare la struttura sintattica di un qualsiasi linguaggio, formale o naturale, grazie a una data grammatica formale.

Nell'ambito della linguistica computazionale questo termine viene utilizzato in riferimento al processo di rappresentazione di una frase o di una qualsiasi stringa di caratteri nei loro *costituenti*, risultante in una struttura dati al albero, che mostra le relazioni sintattiche che intercorrono tra le singole unità delle frasi o parole analizzate.

Il parsing a dipendenze rappresenta uno dei principali framework computazionali di analisi sintattica ad oggi utilizzato, caratterizzandosi principalmente per un'analisi volta a rintracciare le relazioni di dipendenza sintattica tra le parole, dove per ogni relazione viene evidenziato il *dipendente* (dependent o governor) e la sua *testa* (head) sintattica.

L'attività di parsing sintattico viene svolta da un programma detto *parser*, che analizza un flusso continuo di dati (parole) in ingresso, per restituire in uscita un albero sintattico che mostra la struttura sintattica della frase in input, e in cui i nodi rappresentano le parole-unità prese in analisi e gli archi rappresentano le dipendenze grammaticali (connessioni delle *teste* al loro *dipendente*) tra le parole.

Fino ad ora le rappresentazioni sintattiche basate sulle relazioni di dipendenza tra le parole hanno svolto un ruolo piuttosto marginale nella storia della teoria linguistica, come in quella del Natural Language Processing. Negli ultimi anni un incremento di interesse nei confronti degli approcci di questo tipo di parsing conferma che la nozione di dipendenza stia diventando sempre più prominente nella letteratura sul parsing sintattico.

Secondo Covington (2001) il parsing a dipendenze offre a prima vista tre vantaggi rispetto ad altri modelli competitivi:

- le relazioni di dipendenza sono vicine alle relazioni semantiche, necessarie per un successivo stadio di interpretazione linguistica;
- l'albero a dipendenze contiene un nodo per parola. Poiché il lavoro del parser consiste solo nel connettere nodi esistenti e non nel postularne di nuovi, l'attività di parsing è più diretta;
- il parsing a dipendenze si presta ad un'operazione di una parola alla volta, cioè prende in input e analizza le parole una alla volta piuttosto che aspettare per le frasi complete.

A questi vantaggi c'è chi aggiunge che il parsing basato sulle dipendenze permette una trattamento più adeguato delle lingue con un ordine delle parole variabile, dove sono più comuni le costruzioni sintattiche discontinue (Mel'cuk, 1988; Covington, 1990).

In generale le rappresentazioni a dipendenze vengono considerate un compromesso tra l'espressività delle rappresentazioni sintattiche e la complessità del parsing sintattico. Sebbene siano meno espressive di altri modelli, compensano questa mancanza, da una parte, fornendo una codifica relativamente diretta della struttura argomento-predicato, che è rilevante per l'interpretazione semantica, e dall'altra, permettendo una facile disambiguazione. In questo modo, le strutture a dipendenza sono sufficientemente espressive per essere utili in sistemi di natural language processing e allo stesso tempo sufficientemente restrittive per permettere un parsing completo con buoni risultati in termini di accuratezza ed efficienza (Nivre, 2005).

Nonostante i metodi di analisi sintattica basati sul funzionamento di algoritmi di apprendimento automatico presentino diversi limiti, soprattutto a causa della complessità data dalle caratteristiche intrinseche di ambiguità del linguaggio umano, ad oggi essi rappresentano lo “stato dell’arte” nei compiti di annotazione linguistica.

Nei prossimi capitoli verranno introdotti i fondamenti teorici della grammatica a dipendenze e i metodi utilizzati per il parsing a dipendenze.

1.2 La grammatica a dipendenze

Nell'ambito della grammatica normativa, la grammatica a dipendenze si è sviluppata come tipologia di rappresentazione sintattica in Europa, e in particolare nei domini Classici e Slavi (Mel'cuk, 1988). Questa tradizione grammaticale ha raggiunto il livello di formalizzazione più alto con il lavoro di Tensière (1959), solitamente considerato come punto di partenza della moderna tradizione teoretica della grammatica a dipendenze.

Questa tradizione comprende un'ampia ed eterogenea famiglia di teorie e formalismi grammaticali che condividono alcuni presupposti di base sulla struttura sintattica, e in particolare il presupposto che questa consista in una serie di elementi *lessicali,* collegati da relazioni binarie asimmetriche chiamate *dipendenze*. Pertanto, la proprietà formale comune delle strutture a dipendenza è l'assenza di nodi rappresentanti più di un costituente, come nella teoria formale della grammatica a costituenti, in cui la frase viene considerata a partire da combinazioni di sintagmi, a loro volta costituiti da unità più piccole.

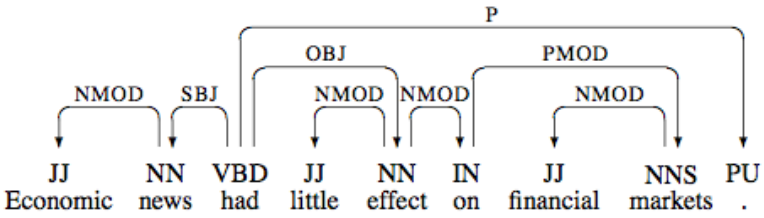


Figura 1.1: Struttura a dipendenze di una frase inglese estratta dalla Penn Treebank

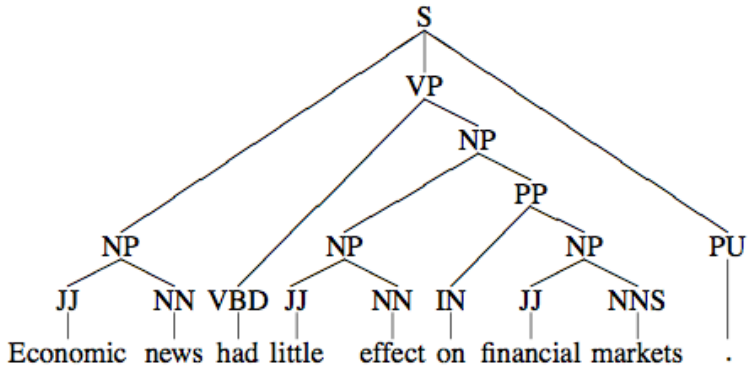


Figura 1.2: Struttura a costituenti di una frase inglese estratta dalla Penn Treebank

La nozione fondamentale di dipendenza viene espressa da Tesnière nei capitoli iniziali di "Elementi di Sintassi strutturale":

> La frase è un ***insieme organizzato*** i cui elementi costitutivi sono le ***parole***. [1.2] Ogni parola, nel momento in cui fa parte di una frase, cessa di essere isolata come avviene nel dizionario. Tra essa e le parole vicine la mente intravede delle ***connessioni***, il cui insieme costituisce la struttura portante della frase. [1.3] Tali connessioni non sono segnalate con alcun mezzo. Ma è indispensabile che esse vengano avvertite dalla mente, altrimenti la frase non sarebbe intelligibile. Quando dico *Alfredo parla*, non intendo dire che «da un lato c'è un uomo che si chiama Alfredo» e dall'altro «che qualcuno parla», ma intendo dire al tempo stesso che «Alfredo compie l'azione di parlare» e che «chi parla è Alfredo». [1.4] Da ciò risulta che una frase del tipo *Alfredo parla* non è composta da due elementi: 1) *Alfredo*, 2) *parla*, bensì da tre elementi: 1) *Alfredo*, 2) *parla* e 3) la connessione che li unisce. [1.5] Le connessioni strutturali stabiliscono tra le parole dei rapporti di ***dipendenza***. In generale, infatti, ogni connessione unisce un termine superiore ad uno inferiore. [2.1] Il termine superiore prende il nome di ***reggente***. Il termine inferiore prende il nome di ***subordinato***. Così nella frase Alfredo parla, parla è il reggente e Alfredo il subordinato. [2.2] Lo studio della frase, che è l'oggetto proprio della sintassi strutturale, è essenzialmente lo studio della sua struttura, che non è altro che la ***gerarchia delle sue connessioni***. [2.6]

La teoria di Tesnière si basa su tre concetti complementari di *connessione*, *unione* e *traslazione*:

- connessione: relazione di dipendenza tra reggente e subordinato

- unione: relazione tra elementi coordinati
- traslazione: relazione tra parole grammaticali o altre parole che cambiano la categoria sintattica di un elemento lessicale.

Per quanto riguarda le relazioni di dipendenza, in letteratura, solitamente, esse vengono dette intercorrere tra una *testa* ("head") e un *dipendente* ("dependent"). Termini alternativi sono *governatore* e *reggente* per la testa e *modificatore* per il dipendente.

I criteri utilizzati per definire le relazioni di dipendenza, e per distinguere la testa e il dipendente all'interno di queste relazioni, sono chiaramente di centrale importanza non solo per la grammatica a dipendenze, ma anche in altri framework in cui la nozione di testa sintattica svolge un ruolo importante.

Alcuni criteri per l'identificazione di una relazione sintattica tra una testa T e un dipendente D in una costruzione C sono stati proposti da (Zwicky, 1985; Hudson, 1990):

1. T determina la categoria sintattica di C e spesso può sostituire C
2. T determina la categoria semantica di C; D da una specifica semantica.
3. T è obbligatoria; D può essere opzionale.
4. T seleziona D e determina se D sia obbligatoria o opzionale.
5. La forma di D dipende da T (accordo or governo).
6. La posizione lineare di D è specificata in riferimento a T.

Questa lista rappresenta un misto di criteri sintattici e semantici, non sempre riconducibili ad una singola coerente nozione di dipendenza. Questo ha portato alcuni studiosi, come Hudson (1990), a suggerire che il concetto di testa abbia una struttura prototipica, per cui le istanze tipiche di questa categoria soddisfano tutti o la maggior parte dei criteri sopraelencati, mentre le istanze più marginali ne soddisfano di meno. Altri autori hanno enfatizzato la necessità di una distinzione tra diversi tipi di relazioni di dipendenza. Secondo Mel'cuk (1988), le parole di una frase possono essere collegate da tre tipi di dipendenze: morfologiche, sintattiche e semantiche. Secondo Nikula (1984), si dovrebbe distingue tra dipendenza sintattica nelle costruzioni endocentriche (dove la testa può sostituire il dipendente senza interrompere la struttura sintattica) ed esocentriche (dove la testa non può essere direttamente sostititutiva del dipendente).

In generale si può dire che la tradizione teorica della grammatica a dipendenze è unita dalla comune assunzione che una parte essenziale della struttura sintattica della frase risieda in relazioni binarie asimmetriche tra elementi lessicali. Sebbene si possa descrivere un nucleo di costruzioni sintattiche per il quale le analisi date da diversi framework teorici concordano sotto molti aspetti, allo stesso

tempo, su diverse questioni, rimangono ancora importanti punti di divergenza. Mentre le strutture testa-complemento e testa-modificatore risultano prestarsi a un'analisi più diretta, ci sono diverse costruzioni su cui non è stato raggiunto un consenso generale in rispetto all'analisi e al tipo di rappresentazione da attribuirgli.
Tra le principali dimensioni di variazione ricordiamo:

- identificazione di ciò che può costituire un nodo in una struttura a dipendenze (parola, lemma, lessema ecc.)
- categorizzazione dei tipi di dipendenza (in generale la maggior parte delle teorie assume un set di funzioni grammaticali più orientate verso la sintassi superficiale, con una maggiore o minore sotto-classificazione elaborata, o un set di ruoli grammaticali orientati semanticamente)
- proprietà formali delle rappresentazioni strutturali
- relazione tra la struttura a dipendenze e l'ordine delle parole (ammissione o meno del vincolo di *proiettività*: un grafo a dipendenze soddisfa il vincolo di proiettività tenendo conto dell'ordine lineare delle parole se, per ogni arco $t \rightarrow d$ e nodo w, w si presenta tra t e d nell'ordine lineare solo se w è dominato da t)
- selezione della testa a livello semantico o sintattico
- trattamento della coordinazione

1.3 Parsing con rappresentazioni a dipendenze

Nelle analisi sintattiche basate sulle rappresentazioni a dipendenze, a causa del grado di formalizzazione relativamente più basso delle teorie di grammatica a dipendenze, le connessioni tra i framework teorici e i sistemi computazionali sono spesso piuttosto indirette. Per questo, generalmente, in questo ambito si parla di parsing con *rappresentazioni* piuttosto che di parsing con *grammatica a dipendenze*.
Nel trattare i diversi sistemi di parsing sintattico a dipendenze si segue solitamente la distinzione effettuata da Carroll (2000) tra due grandi tipi di strategie: l'approccio *grammar-driven* e l'approccio *data-driven*. Questi due approcci possono essere visti come complementari, e molti sistemi esistenti combinano elementi di entrambi. Da una prospettiva analitica, sostanzialmente, si distinguono per quanto riguarda le metodologie adottate nella risoluzione dei problemi presentati dall'analisi di testi di un linguaggio naturale.
Nell'approccio *grammar-driven* viene utilizzata una grammatica formale G per definire a) il

linguaggio L (G) che può essere parsato e b) la classi di analisi che possono essere restituite come output per ogni stringa nel linguaggio.
Nell'approccio *data-driven* una grammatica formale non è più una componente necessaria del sistema di parsing. La mappatura dalla stringa di input alla sua analisi sintattica viene definita da un meccanismo induttivo che si applica su un testo del linguaggio da analizzare. In generale possiamo distinguere tre componenti essenziale in un parser data-driven:

- (1) Un modello formale M che definisce le possibili analisi per le frasi nel linguaggio L
- (2) Un campione di testo S (possibilmente annotato) di L
- (3) Uno schema di inferenza induttivo che definisce le analisi per le frasi di un testo T in L, relativo ad M ed S.

Il funzionamento dei parser data-driven si basa sull'applicazione di algoritmi di apprendimento automatico che a partire da corpora di addestramento annotati con informazione morfo-sintattica e sintattica, costruiscono un modello probabilistico per l'annotazione linguistica del testo di input.
In entrambi i tipi di approcci i criteri che vengono utilizzati per la valutazione di un parser sono quattro:

1. *robustezza*: capacità da parte di un sistema di analizzare qualsiasi frase in input
2. *disambiguazione*: capacità di selezionare l'analisi corretta tra quelle possibili
3. *accuratezza*: precisione o qualità dell'analisi linguistica assegnata alla frase di un testo dal parser
4. *efficienza*: capacità di analisi linguistica con il minimo impiego di risorse e di tempo

Nell'ambito del parsing a dipendenze non si può dire che uno dei due approcci abbia dimostrato risultati migliori rispetto all'altro. Si può dire che i requisiti contrastanti di robustezza, disambiguazione, precisione ed efficienza danno luogo ad un problema di ottimizzazione complesso che richiede sempre una ottimizzazione congiunta. In quanche modo, l'ampia varietà di metodi d'analisi può considerarsi il risultato di diverse strategie di ottimizzazione e obiettivi diversi.
L'approccio grammar-driven, nella sua forma più pura, parte da un sistema con un'accuratezza ottimale, nel senso che vengono considerate solo le frasi di cui può essere derivata l'analisi corretta, e gradualmente si propone di migliorare il sistema rispetto alla robustezza e la disambiguazione. Tuttavia, questo tipo di sviluppo può compromettere l'efficienza, che deve quindi essere ottimizzata insieme a robustezza e disambiguazione. Al contrario, l'approccio data-driven, nella sua forma più radicale, parte da un sistema con robustezza e disambiguazione ottimali, nel senso che ogni frase ottiene esattamente un'analisi, e gradualmente cerca di migliorare il sistema in termini di

accuratezza. Ancora una volta, questo può portare a problemi di efficienza, che deve essere ottimizzata insieme all'accuratezza.

Negli ultimi anni lo sviluppo di programmi di analisi sintattica si è orientato nella direzione dei sistemi data-driven rendendo sempre più necessarie la disponibilità di dati annotati manualmente e schemi di annotazione di riferimento per l'addestramento dei programmi di parsing. Nel prossimo capitolo esamineremo nel particolare il progetto entro il quale è stata definita la treebank di dati linguistici utilizzata nel presente lavoro.

2 Universal Dependencies

In questo capitolo verrà presentata lo schema di annotazione e il corpus (treebank) di dati linguistici utilizzata nel presente lavoro, ovvero quello delle Universal Dependencies. Questa risorsa rappresenta una delle ultime iniziative di annotazione linguistica avviate a livello internazionale per la definizione di risorse e schemi di annotazione da poter utilizzare come standard di riferimento multilingue.

2.1 Universal Dependencies

Il progetto *Universal Dependencies* nasce al fine di creare un valido modello di annotazione grammaticale di tipo inter-linguistico e di fornirne le linee guida per una sua applicazione interlinguistica.

Basandosi su diverse iniziative precedenti, tra cui *Inteset*, *Google Universal Part-of-Speech Tags*, *HambleDT*, *Universal Dependency Treebanks* e *Universal Stanford Dependencies*, UD si propone di definire una grammatica universale che sia in grado, a partire da un set finito e ridotto di categorie grammaticali dalla valenza potenzialmente universale, di rappresentare a livello strutturale (*morfologico* e *sintattico*) diverse tipologie linguistiche, in modo indicativo e utile per applicazioni di Natural Language Processing.

Il modello proposto punta ad essere di riferimento come standard di annotazione grammaticale, poiché non solo si basa su uno schema consistente, chiaro e semplice, ma, mettendo in evidenza e riassumendo le similarità interlinguistiche in modo proficuo per possibili interazioni interlinguistiche, attenua i principali problemi di incompatibilità e rigidezza che ricercatori e sviluppatori riscontrano nella progettazione di programmi di parsing accurati ed efficienti e in diversi altri ambiti di applicazione di NLP . La selezione di un set di categorie morfologiche e sintattiche che fossero riconducibili al maggior numero di lingue e che inglobassero le parti del discorso più frequenti è stato fondamentale per massimizzare i parallelismi tra gli schemi di annotazione di lingue molto diverse, anche a livello tipologico.

L'ampia varietà di schemi d'annotazione disponibile, data dall'interferenza di diverse definizioni teoriche tra i costruttori di banche di dati, sarebbe preferibile da superare per diverse ragioni:

- nelle applicazioni multilingue che coinvolgono l'uso di parser, l'assenza di uniformità tra le

diverse rappresentazioni date in output dai parser richiede interfacce specializzate per ogni lingua;

- l'uso di standard di annotazione inconsistenti, oltre ad ostacolare l'apprendimento interlinguistico, rende impossibile la valutazione delle performance dei parser nei programmi di evaluation, poiché non si può distinguere tra gli errori e le discrepanze tra gli standard;
- nel parsing statistico, risulta difficile realizzare un uso preciso e inappuntabile della conoscenza linguistica poiché non è possibile una rappresentazione consistente delle categorie e delle strutture linguistiche tra le diverse lingue.

UD risponde non si propone come teoria linguistica né come modello di parsing ottimale. Risponde in primo luogo a un piano di genere pratico, volto alla realizzazione di una risorsa polifunzionale, applicabile per facilitare lo sviluppo di parser multilingue, l'apprendimento automatico interlinguistico e la ricerca su programmi di parsing da un punto di vista tipologico. Questa risorsa è inoltre un utile strumento per la costruzione e la valutazione di taggers non supervisionati e interlinguistici, per un'osservazione comparativa dell'accuratezza di taggers supervisionati e, infine, per l'addestramento di POS taggers con tagset comuni tra diverse lingue.

I dataset attualmente disponibili, appartenenti a ben 22 domini linguistici differenti, tra cui quello dell'italiano, sono stati realizzati attraverso programmi di mapping tra gli schemi d'annotazione originari di treebank preesistenti e il tagset di UD.

Per incrementare l'utilità della risorsa, oltre ad all'ampliamento della copertura delle lingue e della diversità tipologica, attualmente, una delle principali sfide poste dalla ricerca è quella di migliorare la qualità delle treebank da convertire negli standard UD, con particolare riferimento alla reale consistenza intelinguistica.

2.2 Il tagset UD: principi di progettazione

La struttura dello schema di annotazione adottato in UD si basa su una segmentazione delle *frasi* in *parole*, dove la parola è descritta da una serie di *proprietà morfologiche* e interconnessa in un sistema di dipendenze tramite *relazioni sintattiche*. L'unità di base di analisi è dunque la parola, intesa sintatticamente (non a livello fonologico o ortografico), secondo un modello in linea con le ipotesi lessicaliste e la morfologia lessicale.

2.2.1 Tokenizzazione

La segmentazione in parole avviene secondo regole specifiche predefinite per ogni lingua: l'idea di base è che ad ogni unità sintattica corrisponda un'unica descrizione morfologica con un unico lemma, un part-of-speech tag e un set di tratti morfologici, e che questa unità sia connessa alle altre parole della frase tramite una singola funzione sintattica. Poiché i principi di segmentazione variano notevolmente a seconda della tipologia e della distribuzione specifiche della lingua, per ogni caso viene fornita una specifica documentazione che descrive come vengono individuate e trattate le unità di analisi, con particolare attenzione per le locuzioni e i tipi di tokens composti da più *parole sintattiche* (in italiano un esempio è fornito dai clitici che, in alcuni casi si attaccano a un morfema di una certa parte del discordo pur svolgendo una funzione sintattica indipendente). L'annotazione morfologica e sintattica viene definita a livello di parola ma può essere fornito un mapping euristico verso il livello di token.

2.2.2 Annotazione morfologica

La descrizione morfologica di una parola *sintattica* (tenendo in considerazione all'interno della frase la sua funzione sintattica) consiste in tre livelli di rappresentazione:

- un *lemma* corrispondente al contenuto semantico della parola
- un *part-of-speech tag* corrispondente alla categoria lessicale astratta della parola
- un insieme di *tratti linguistici* (“features”) corrispondenti alle proprietà lessicali e grammaticali associate al lemma o alla specifica forma della parola.

I lemmi I lemmi sono generalmente determinati da dizionari e lessici specifici per ogni lingua e il loro campo contiene generalmente la forma canonica o base della parole, corrispondente a quella trovata nei dizionari. Se il lemma non è disponibile, la sua assenza viene indicata da un underscore (“_”).
I part-of-speech tag e le proprietà grammaticali, contrariamente ai lemmi, vengono estratti da un inventario universale, predefinito.

I part-of-speech tag L'insieme dei part-of-speech tag universali è definito in una lista contenente 17 tag, basata sul *Google Universal Part-of-speech Tagset*, a sua volta rielaborazione dei tagset utilizzati nella CoNLL-X shared task. I tag universali devono essere utilizzati in tutte le banche di dati costruite sul modello delle Universal Dependencies. E' possibile che alcuni tag non vengano usati in alcune lingue. Nei casi più specifici, analisi più accurate vengono riportate nel campo dei tratti linguistici.

Inoltre il formato ConNLL-U permette l'aggiunta di un ulteriore campo “POSTAG”, preso da un tagset specifico di una lingua o di un corpus.
I POS tag universali sono formati solo da lettere maiuscole dell'alfabeto latino [A-Z]. A ogni parola viene associato solo un tag, nel caso in cui nessuno dovesse essere appropriato si utilizza il tag “X”.

Tabella 2.2.1: Part-of-speech tag universali

Classe aperta di parole	Classe chiusa di parole	Altro
ADJ: aggettivo	ADP: apposizione	PUNCT: punteggiatura
ADV: avverbio	AUX: verbo ausiliare	SYM: simbolo
INTJ: interiezione	CONJ: congiunzione coord.	X: altro
NOUN: nome	DET: determinante	
PROPN: nome proprio	NUM: numerale	
VERB: verbo	PART: particella	
	PRON: pronome	
	SCONJ: congiunzione subordinante	

Le *features* Il campo dei tratti (“features”) fornisce ulteriori informazioni sulla parola, la sua part-of-speech e le sue proprietà morfosintattiche. Ogni tratto ha la forma Nome=Valore e ad ogni parola può essere associato un numero variabile di tratti, separati tra di loro dalla barra verticale (“ | ”). All'utenza viene data la possibilità di estendere il set di tratti universali standardizzati con tratti specifici per la lingua, se necessario. I tratti aggiuntivi devono essere documentati e devono seguire i principi generali di formato.

Tabella 2.2.2: Tratti linguistici universali (“features”)

Tratti lessicali	Tratti flessivi	
PronType: tipo pronominale	*Nominali*	*Verbali*
NumType: tipo numerale	Gender: genere	VerbForm: forma verbale
Poss: possessivo	Animacy	Mood: modo
Reflex: riflessivo	Number: numero	Tense: tempo
	Case: caso	Aspect: aspetto
	Definite	Voice: diatesi
	Degree: grado	Person: persona
		Negative: negabilità

Tutti gli identificatori (di tipo nome o valore) sono formati da lettere latine o occasionalmente da cifre numeriche.
In alcuni casi, quando lo stesso tratto è marcato più di una volta per singola parola, è possibile stratificare il tratto stesso aggiungendo tra parentesi quadre un ulteriore identificatore per indicare e distinguere lo strato considerato.

2.2.3 Annotazione sintattica

L'annotazione sintattica in UD consiste in un sistema gerarchico valenziale di relazioni di dipendenza tra le parole, con una speciale relazione "*root"* (radice) per quelle che non dipendono da nessun'altra parola.
Ogni frase è associata ad un set di dipendenze base che formano un grafo diretto aciclico, ovvero un albero radicato, che rappresenta la struttura sintattica della frase. In aggiunta, alcune costruzioni sintattiche possono introdurre ulteriori dipendenze che possono essere rappresentate nella versione *potenziata* ("enhanced") delle UD, dove vengono codificate nel campo DEPS del formato CoNLL-U. Tuttavia non sono state ancora sviluppate dettagliate linee guida per questo tipo di rappresentazione.
Le relazioni di dipendenza selezionate all'interno delle UD sono state pensate con lo scopo di definire un insieme di funzioni grammaticali trasversali al maggior numero di lingue, massimizzando i parallelismi di tipo sintattico.
Il principio di base prevede che le relazioni di dipendenza intercorrano principalmente tra parole piene, senza la mediazione delle parole grammaticali (o vuote), le quali vengono trattate come dirette dipendenti della parola piena a cui sono più strettamente connesse. La punteggiatura viene connessa alla testa della frase o della proposizione. Nell'albero risultante i nodi interni saranno rappresentati da parole piene, mentre le parole grammaticali, in quanto non modificate da alcun dipendente, e dunque non raggruppate in una struttura annidata, rappresenteranno le foglie.
Si distinguono quattro casi in cui le parole grammaticali eccezionalmente fanno da testa ad altre componenti della frase:

- Locuzioni propositive o congiuntive: le parole che formano la locuzione sono connesse tra di loro da una speciale relazione di dipendenza "*mwe"*. Quando la locuzione svolge un ruolo funzionale il suo primo componente apparirà superficialmente come una parola grammaticale con dipendenti.
- Parole grammaticali coordinate: così come avviene per le parole piene, il primo elemento coordinato viene trattato come testa della congiunzione e degli altri elementi coordinati.

- Modificatori: una ristretta classe di modificatori, come elementi di negazione (relazione "*neg*") e avverbi (relazione "*advmod*" e "*nmod*"), possono dipendere da parole grammaticali.
- Promozione attraverso l'elisione della testa: quando la testa di una parola piena è elisa, quest'ultima viene "promossa" a svolgere la funzione normalmente assunta dalla testa assente.

I tre tipi di espressione annotati in una struttura testa-dipendente, dove tutti gli elementi non iniziali dipendono dal primo, e dove solo il primo elemento può avere dipendenti sono le locuzioni fisse, i nomi, e le frasi in lingua straniera.

Sistema di classificazione Le relazioni sintattiche prese in considerazione, riscritte sul modello delle Universal Stanford Dependencies (de Marneffe et al. 2014), sono classificate secondo un principio di differenziazione strutturale che distingue frasi di tipo nominale da frasi di tipo predicativo e da un'altri tipi di modificatori. La classificazione distingue inoltre gli "argomenti centrali" (soggetti, oggetti, elementi complementari) da gli altri tipi di dipendenti. Oltre alle relazioni di dipendenza universali di base è sempre possibile definire e aggiungere sottotipi di relazioni specifici per costruzioni di particolare rilevanza in una data lingua.

Diatesi Per quanto riguarda la funzione di soggetto, viene fatta distinzione tra diatesi attiva ("*nsubj*", "*csubj*") e diatesi passiva ("*nsubjpass*", "*cnsubjpass*"). Nelle proposizioni subordinate si fa distinzione tra proposizioni con controllo obbligatorio ("*xcomp*") da proposizioni con soggetto indipendente, mentre non si fa tra proposizioni dirette e indirette.

Coordinazione La coordinazione viene trattata asimmetricamente: la testa della relazione è il primo elemento coordinante e tutti gli altri elementi ne sono dipendenti attraverso la relazione *conj*. Gli elementi coordinanti e la punteggiatura delimitanti vengono connessi usando rispettivamente, le relazioni cc e punct. I token connessi tramite la relazione punct non possono avere dipendenti e dipendono solo da parole piene.

Punteggiatura I token con la relazione "*punct*" si attaccano sempre alle parole piene e non possono mai avere dipendenti:

- un segno di punteggiatura che separa unità coordinate si attacca al primo elemento coordinato

- un segno di punteggiatura che precede o segue un'unità subordinata si attacca a questa unità

Tabella 2.2.3: Tabella delle relazioni di dipendenza fornita sul sito ufficiale di UD

Core dependents of clausal predicates			**Non core dependents of clausal predicates**		
Nominal dep	*Predicate dep*		*Nominal dep*	*Predicate dep*	*Modifier word*
nsubj	csubj		nmod	advcl	admod
dobj	ccomp	xcomp			
iojb					

Noun dependents			**Special clausal dependents**		
Nominal dep	*Predicate dep*	*Modifier word*	*Nominal dep*	*Auxiliary*	*Other*
nummod	acl	amod	vocative	vocative	mark
appos		det	discourse	auxpass	punct
nmod		neg	expl	cop	

Coumpounding and unanalyzed			**Loose joining relations**		
compound	mwe	goeswith	list	parataxis	remnant
name	foreing		dislocated		reparandum

Case-marking, prepositions, possessive	**Coordination**		
case	conj	cc	punct

Other	
Sentence head	*Unspecified dependency*
root	dep

Per un approfondimento delle funzioni delle singole relazioni di dipendenza si rimanda all'Appendice.

2.2.4 Il formato CoNLL-U

Il formato utilizzato nelle UD è una versione revisionata del formato CoNLL-X. Le annotazioni sono codificate in file plain text (UTF-8) con tre tipi di righi:

- righi di parola, contenenti l'annotazione di una parola/token suddivisa in 10 campi, separati da una singola tabulazione

- righi vuoti indicanti i confini di parola
- righi di commento introdotti dall'hash (#)

Campi di classificazione

1. ID: indice di parola, numero intero pari a 1 all'inizio di ogni frase
2. FORM: forma della parola o simbolo di punteggiatura
3. LEMMA: lemma o radice della parola
4. CPOSTAG: part-of-speech tag universale
5. POSTAG: part-of-speech tag specifico della lingua in analisi
6. FEATS: lista di tratti morfologici
7. HEAD: testa del token corrente, che può essere il valore dell'ID o lo zero (0)
8. DEPREL: relazione di dipendenza del token corrente con l'HEAD (root se e solo se la Head è uguale a zero) o un particolare sottotipo specifico della lingua in analisi
9. DEPS: lista di dipendenze secondarie
10. MISC: qualsiasi annotazione eventuale.

Es. Frase nel formato CoNLL-U.

1	Vorrei	volere	V	VM	Mood=Cnd|Number=Sing|Person=1|Tense=Pres					
VerbForm=Fin		3	aux	_	_					
2	ora	ora	B	B	_	3	advmod	_	_	
3	entrare	entrare	V	V	VerbForm=Inf	0	root	_	_	
4	brevemente		brevemente	B	B	_	3	advmod	_	_
5	nel	in	E	EA	Gender=Masc|Number=Sing		6	case	_	_
6	merito	merito	S	S	Gender=Masc|Number=Sing		3	nmod	_	_
7	.	.	F	FS	_	3	punct	_	_	

2.3 La IUDT (Italian Universal Dependency Treebank)

La versione della treebank italiana annotata secondo lo schema delle Universal Dependencies è stato il risultato di diverse operazioni di *merging* e di conversione. La disponibilità limitata di risorse di addestramento, data dai costi ingenti necessari allo sviluppo di una grande banca di dati, ha contribuito, da una parte, a scoraggiarne la creazione ex novo, dall'altra a concentrarsi sul riutilizzo di datasets già esistenti.

Nel quadro generale di ricerca nazionale, tentativi preliminari di costruzione di una dependency treebank a cui ci si potesse riferire come standard nell'ambito del dependency parsing, risalgono ai primi esperimenti di merging di corpora già precedentemente assemblati per la lingua italiana.

Nello specifico, la TUT, la Turin University Treebank (Bosco et al. 2000), e la ISST-TANL, inizialmente rilasciata come la ISST-CoNLL per la CoNLL-2007 shared task (Montemagni, Simi 2007), sono state le principali risorse per la costruzione, prima della MIDT (la Merged Italian Dependency Treebank), ottenuta tramite la fusione dei due diversi corpora, e poi della ISDT (la Italian Stanford Dependency Treebank), unica treebank per l'italiano annotata secondo lo schema delle Stanford Dependencies, rilasciata in occasione del dependency parsing shared task di Evalita-2014 (Bosco et al. 2014) e poi utilizzata come punto di partenza per la definizione, tramite conversione, di IUDT, il corpus annotato secondo il modello delle Universal Dependencies. La prima versione di IUDT è stata pubblicata nel gennaio del 2015.

Tabella 2.3.1: Composizione di IUDT.

Formato orginale	Fonte	Genere	Dimensione in tokens	Dimensione delle frasi
TUT-CONLL	Evalita 2011 Dependency parsing	Testi legali, articoli di giornale, articoli di Wikipedia	116,986	3,842
ISST-TANL	Evalita 2011 Domain adaptation task	Articoli di giornale	93,721	4,135
ISST-TANL	SPLeT 2012	Testi legali: direttive europee	7,200	260
MIDT	Several QA competitions	Domande	26,078	2,228
MIDT	Evalita 2014 Dependency parsing:test data set (parziale)	Articoli di giornale	8,776	304
TUT-CONLL	Parallel TUT (parte italiana)	Generi vari	63,899	2,129
		Totale	316,660	12,88

Le profonde differenze tra le risorse disponibili, divergenti per quanto riguarda sia le varietà di composizione interna (tipologie di testi selezionati), sia le dimensioni, sia gli schemi di annotazione adottati, ha reso necessaria un'operazione di standardizzazione per migliorarne non solo l'applicabilità ma anche l'affidabilità e a livello inter-linguistico, la comparabilità con altre risorse.

I processi di merging e di conversione hanno contribuito a indirizzare la ricerca verso il consolidamento di un framework teorico ottimale in riferimento alle peculiarità linguistiche, e in particolare morfo-sintattiche, dell'italiano.

Se da un punto di vista di formattazione, già da tempo, la maggior parte delle treebank utilizza il formato di codifica del CoNLL come standard di riferimento, gli schemi di annotazione grammaticale delle treebank sopracitate cambiano in maniera incisiva in riguardo a diverse dimensioni di variazione, che vanno dai criteri di selezione della testa, dalla granularità, e dunque, dal livello di dettaglio dei tagset delle relazioni di dipendenza, fino ai criteri generali di annotazione.

Le varie soluzioni intermedie sono state raggiunte attraverso euristiche di tipo comparativo tra i vari sistemi di annotazione, e soprattutto attraverso la valutazione in accuratezza delle performance eseguite da programmi di parsing su risorse annotate diversamente. I risultati ottenuti hanno contribuito a fare da guida nelle operazioni di merging, fornendo una base solida nei processi di conversione e adattamento delle risorse verso gli standard internazionali.

Attraverso le principali dimensioni di variazione degli schemi di annotazione, le analisi si sono concentrate sulle categorie di contenuto, vagliando somiglianze e differenze fino a identificare un insieme di costruzioni sintattiche che potessero fungere da ponte tra le annotazioni.

La metodologia utilizzata adottata per armonizzare e fondere i diversi schemi di annotazione ha seguito queste fondamentali passaggi:

- un'analisi comparativa tra gli schemi di annotazione delle risorse di partenza e quelli di target, con particolare attenzione per i principali punti di differenza;
- l'analisi della performance di parser a dipendenze aggiornati e affidabili utilizzando per l'addestramento le treebank preesistenti e quelle di target;
- un collegamento tra gli schemi di annotazione delle risorse di partenza verso un unico set di categorie di dati.

La discussione dei processi di armonizzazione e di conversione che hanno portato alla creazione di MIDT e di ISDT si trova in (Bosco, Montemagni, Simi, 2012) e in (Bosco, Montemagni, Simi, 2013 e 2014).

2.3.1 Schema di annotazione di IUDT

Lo schema di annotazione adottato coincide con quello di UD, eccetto qualche differenza riguardante i tratti linguistici e le relazioni di dipendenza. In particolare nella treebank italiana IUDT non vengono utilizzate:

- come features: *Animacy*, *Aspect*, *Voice*
- come relazioni di dipendenza: *goeswith*, *list*, *dislocated*, *remnant*, *reparandum*

Per quanto riguarda le POS, rispetto allo schema specifico per l'italiano (ISDT) rimangono invariate le categorie INTJ, CONJ, NUM, PART, SCONJ, PUNCT e SYM. Vengono invece modificati nome e classificazione delle POS elencate nella tabella 2.3.2.

Tabella 2.3.2: Mappatura tra le POS di IUDT e le POS specifiche dell'italiano

Parts-of-Speech tags di IUDT	Parts-of-Speech tags corrispondenti specifiche dell'italiano
ADJ: adjective	A: Adjective NO: Ordinal Number
ADV: adverb	B: Adverb BN: Negation adverb
ADP: adposition	E: Preposition
AUX: auxiliary verb	VA: Auxiliary verb VM: Modal verb
DET: determiner	DET: determiner RD: Definite article RI: Indefinite article DE: Exclamative determiner DI: Indefinite determiner DQ: Interrogative determiner DR: Relative determiner DD: Demonstrative determiner T: Predeterminer AP: Possessive adjective
N: noun	S: Common noun
PRON: pronoun	PC (clitic pronouns) → PronType=Clit PD (demonstrative pronouns) →PronType=Dem PE (personal pronouns) → PronType=Prs PI (indefinite pronouns) → PronType=Ind PP (possessive pronouns) → PronType=Prs | Poss=Yes PQ (interrogative pronouns) → PronType=Int PR (relative pronouns) → PronType=Rel
PROPN: propen noun	SP: Proper noun
VERB: verb	V: Main verb

3 Risorse, Strumenti e Metodologie d'analisi

Le analisi linguistiche del presente studio sono state predisposte attraverso una precisa metodologia di indagine che ha richiesto l'uso congiunto di diverse risorse e strumenti di applicazione linguistico-computazionali.

La selezione dei dati e delle tecnologie di analisi è avvenuta rivolgendo particolare attenzione al loro grado di affidabilità e accuratezza descritto in letteratura, in linea con gli standard dello *stato dell'arte* nell'ambito del dependency parsing.

L'elaborazione dei dati d'analisi è stata realizzata attraverso una metodologia a tre fasi:

- **Fase 1**: annotazione linguistica automatica di un grande corpus di testi giornalistici italiani in una procedura a tre fasi (*sentence splitting*, *POS tagging* e *syntactic parsing*). Gli strumenti ultilizzati per l'analisi sono stati addestrati sulla treebank *gold* di IUDT descritta nel precedente capitolo;
- **Fase 2**: creazione di un modello statistico tramite l'applicazione di un algoritmo (LISCA) che sfrutta un preciso set di *feature*s linguistiche estratte probabilisticamente dal corpus automaticamente parsato nella fase 1;
- **Fase 3**: il calcolo di un punteggio di prototipicità (o plausibilità) per ogni relazione di dipendenza presente all'interno della treebank *gold* (utilizzata nella fase 1 di addestramento) utilizzando il modello statistico creato a partire dal corpus giornalistico durante la fase 2.

I punteggi di plausibilità ottenuti in fase 3 sono stati utilizzati per ordinare le relazioni di dipendenza osservate nella IUDT. Tali punteggi hanno costituito il focus delle osservazioni linguistico-computazionali riportate nel capitolo 4.

Nei prossimi paragrafi verranno descritte nello specifico tutte le risorse utilizzate e le fasi di analisi presentate.

3.1 I dati linguistici

La treebank *gold* utilizzata nelle fasi di preparazione dei dati è stata la IUDT, selezionata in quanto standard di riferimento nel dominio delle banche di dati linguistici per la lingua italiana.

Nella prima fase di parsing la IUDT è stata presa come corpus di addestramento per il parser statistico DeSR. Successivamente, nell'ultima fase di analisi, è stata utilizzata per l'elaborazione dei

dati finali, tramite una funzione di ordinamento della plausibilità eseguita dall'algoritmo di LISCA in relazione alle relazioni di dipendenza interne alla treebank.
Il campione di testi che è stato annotato automaticamente per la formulazione del modello statistico di LISCA è stato un corpus di 1,104,237 frasi (22,830,739 parole–tokens), composto da articoli di varietà giornalistica pubblicati sul giornale italiano la Repubblica, estratti a loro volta dal CLIC-ILC Corpus (Marinelli e altri, 2003).

3.2 Strumenti di annotazione linguistica automatica

Il campione utilizzato a partire dal quale LISCA ha creato il suo modello statistico, è stato prima trattato, in tre fasi di analisi computazionale:

- *Sentence splitting*: individua le frasi che compongono il corpus;
- *Part-of-speech tagging*: identifica la struttura linguistica in modo incrementale secondo i processi di tokenizzazione, lemmatizzazione e analisi morfo-sintattica, attribuendo a ciascun *token* il lemma corrispondente e la categoria morfo-sintattica di appartenenza;
- *Syntactic parsing*: analizza la struttura sintattica della frase identificando le relazioni di dipendenza che connettono le singole unità (token o parole) alla loro testa sintattica, specificando il tipo di relazione.

3.2.1 ILC-POS-Tagger

Per il p*art-of-speech tagging* è stato utilizzato il sistema ILC-POS-Tagger (Dell'Orletta, 2009) che ha un'accuratezza del 96,34% nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati.

3.2.2 DeSR: *Dependency Shift Reduce*

L'analisi e la rappresentazione delle strutture sintattiche del corpus-campione sono state generate automaticamente dal programma di parsing statistico a dipendenze DeSR (Attardi, 2006), addestrato precedentemente sulla treebank gold costruita per l'italiano secondo lo schema di UD.
Desr è un parser a dipendenze di tipo deterministico-incrementale che utilizza una variante dell'approccio di (Yamada and Matsumoto, 2003).
Si distingue per i seguenti tratti:

- accuratezza: è vicino agli standard dello state of the art;

- efficienza: può parsare fino a 200 frasi al secondo;
- multilingue: a partire da un corpus annotato può essere addestrato su diverse lingue
- personalizzabile: le features utilizzate nell'addestramento possono essere personalizzate.

Tecnica Il funzionamento del parser si basa su un algoritmo di tipo bottom-up in stile Shift/Reduce che analizza le frasi di input scansionandole in ordine, da sinistra verso destra o viceversa, stabilendo ad ogni passo l'opportuna azione di parsing attraverso un *classificatore* basato su un set di *features* rappresentative dello stato di parse corrente.
Le regole di parsing sono speciali e permettono di essere applicate in modo deterministico in un singolo passaggio. Inoltre attraverso un file di configurazione è possibile specificare il set di features da usare (per es. POS tag, lemma, tratti morfologici). Il parser può essere configurato scegliendo tra diversi algoritmi di apprendimento (*Averaged Perceptron*, *Maximum Entropy*, *memory-based learning* con *TiMBL*, *support vector machines* con *libSVM*), fornendo modelli di features personalizzati dall'utente, e selezionando i formati di input e di output (incluso il formato CoNLL).
In (Attardi *e altri*, 2009) DeSR ha mostrato buone percentuali di affidabilità, risultando affidabile per l'87,71% secondo la metrica *UAS* (valutazione della relazione di dipendenza tra parole della frase) e per l'83,38% secondo la metrica *LAS (*identificazione simultanea del tipo di dipendenza e della testa sintattica) (Montemagni, 2013).

3.3 LISCA: *Linguistically-driven Selection of Correct Arcs for Dependency Parsing*

LISCA (Dell'Orletta *e altri*, 2013) è uno strumento di valutazione che viene applicato nell'ambito del dependency parcing al fine di migliorare le performance di parser a dipendenze di tipo data-driven.
Se da una parte la creazione manuale e supervisionata di dati di addestramento rappresenti il metodo più sicuro per ottenere una maggiore accuratezza da parte dei parser, o anche per meglio adattarli a diversi domini, a seconda dei dati di addestramento originari, dall'altra parte questo tipo di pratica è altamente costoso, sia in termini di tempo, sia in termini di sforzi umani necessari. Per questo motivo l'attenzione dei ricercatori si è orientata verso nuove tecniche e metodologie di potenziamento che prevedono l'applicazione di strumenti come LISCA che valutano le performance dei parser statistici calcolando la *plausibilità* dei loro output.
L'accuratezza dei dati annotati automaticamente viene così stimata a partire dalle costruzioni

correttamente identificate dal parser, il che vuol dire, a partire dagli archi o relazioni di dipendenza correttamente parsati.

Questa tecnica viene eseguita con l'obbiettivo di:

- migliorare l'accuratezza dei parser statistici
- adattare i parser statistici a corpora di domini diversi rispetto a quelli su cui vengono addestrati
- supportare la creazione di treebank manualmente annotate, identificando le aree problematiche per il pre- e post-processing umano.

L'algoritmo di LISCA si inserisce in un ambito di ricerca nel quale si distinguono, da un punto di vista metodologico, tre classi di approcci di ricerca, a seconda che nell'algoritmo di rilevazione dell'arco corretto vengano usati set di regole, dati di addestramento o ampie quantità di dati automaticamente analizzati:

- approcci *basati su regole*: metodi di rilevamento di errore basati su una grammatica "*gold*", cioè attraverso la comparazione di regole della grammatica *gold* con le regole indotte dai dati automaticamente parsati, con il fine ultimo di identificare anomalie nelle treebank
- approcci *supervisionati*: metodi che usano statistiche estratte da dati *gold* per rilevare variazioni nelle treebank che potrebbero rappresentare potenziali errori attraverso lo sviluppo di classificatori di tipo comparativo per il rilevamento di errori o per l'identificazione di analisi corrette
- approcci *non supervisionati*: metodi che applicano il punteggio dell'*Informazione mutua puntuale* (o una funzione ad essa vicina) a specifiche configurazioni sintattico-lessicali, usando statistiche estratte da un'ampia quantità di dati automaticamente analizzati per calcolare l'affidabilità degli archi, con finalità di auto-training.

LISCA segue un approccio di tipo non supervisionato finalizzato all'assegnazione di un punteggio di qualità ad ogni arco generato all'interno dell'output di un parser a dipendenze.

Lo scopo è quello di produrre un ranking decrescente di archi che ordini quest'ultimi per grado di correttezza, dal più corretto al meno corretto. Per fare ciò vengono sfruttate le statistiche estratte da un ampio corpus su un set di tratti strutturali linguisticamente motivati e basati su una struttura a dipendenze. In questo caso il corpus giornalistico

Queste statistiche vengono utilizzate per assegnare un punteggio di qualità ad ogni arco di ogni frase analizzata sintatticamente appartenente al dominio del corpus parsato automaticamente.

In (Dell'Orletta *e altri*, 2013) LISCA è stato testato con successo su due dataset appartenenti a due differenti domini e in tutti gli esperimenti ha superato in prestazioni diversi modelli standard,

dimostrando così di essere in grado di rilevare archi corretti anche rappresentando peculiarità dipendenti dal dominio.

3.3.1 L'algoritmo di LISCA

LISCA prende come input un set di frasi parsate automaticamente e assegna ad ogni arco di dipendenza un punteggio che quantifica la sua affidabilità, dove un arco di dipendenza *a* è definito come una tripla (*d*, *h*, *t*) dove *d* è l'unità (parola-token) *dipendente*, *h* è la testa sintattica o il governatore, e *t* è il tipo di relazione di dipendenza che connette *d* ad *h*. Il punteggio qualitativo assegnato all'arco viene poi utilizzato per ordinare gli archi di dipendenza in ordine di affidabilità.
L'algoritmo opera in due passi:

- step 1: colleziona statistiche su un set di tratti linguisticamente motivati, estratti da un ampio corpus di frasi parsate;
- step 2: calcola un punteggio qualitativo per ogni collegamento di dipendenza di una frase appena parsata utilizzando le statistiche estratte dal corpus durante lo step 1

Selezione dei tratti (o *features*) I tratti linguistici che stanno alla base di LISCA sono tutti finalizzati al caratterizzare l'arco che viene analizzato, tenendo conto delle proprietà strutturali (sia globali che locali) dell'albero a dipendenze che lo include.
In particolare, un primo set di tratti è finalizzano a posizionare l'arco di dipendenza all'interno dell'albero, tenendo conto sia della sua struttura gerarchica sia dell'ordine lineare delle parole. Questo set di tratti, basato sulla struttura globale dell'albero, è accompagnato da tratti locali rappresentati dalla lunghezza della relazione di dipendenza, dalla direzione e dalla plausibilità.
I tratti selezionati vengono definiti "linguisticamente motivati" in quanto basati sulla struttura dell'albero a dipendenze, e in particolare in quanto focalizzati su strutture linguistiche che in letteratura si è ampiamente concordi nel riconoscere riflettano la complessità delle frasi a livello sintattico e a livello di parsing.

Localizzazione di un arco di dipendenza all'interno della struttura globale dell'albero Questo set di tratti è finalizzato a localizzare un dato collegamento (o link) di dipendenza all'interno della struttura globale della frase. Questo viene fatto focalizzandosi sul dipendente *d* dell'arco *a* considerato. Un primo tratto complesso analizzato per la localizzazione di *a* all'interno della struttura gerarchica dell'albero, tiene conto di tre fattori:

- della distanza di *d* dal nodo radice

- della distanza di *d* dal nodo foglia più vicino
- della distanza di *d* dal nodo foglia più lontano

Per questo specifico fine i cammini di dipendenza che includono *a* e che vanno dal nodo radice fino ai nodi foglia di *d* più vicini e più lontani vengono rispettivamente ricostruiti. In entrambi i casi, viene selezionato il cammino di dipendenza più corto, la cui lunghezza viene misurata contando il numero di nodi attraversati. La lunghezza di questi due cammini di dipendenza viene usata per caratterizzare il posizionamento dell'arco *a* all'interno della struttura a dipendenze.

A questo tratto globale si aggiungono altri due tratti che si focalizzano su sotto-alberi a dipendenze locali, finalizzati alla localizzazione di *d* tenendo in considerazione l'ordine lineare di superficie delle parole:

- Il primo tratto fa riferimento al sotto-albero dipendente da *d* e conta tutti i suoi immediati dipendenti (o figli rispetto alla struttura ad albero) che vengono divisi in due classi a seconda precedano (pre-dipendenti) o seguano (post-dipendenti) la testa *d* nell'ordine lineare delle parole all'interno della frase
- Il secondo tratto fa riferimento ai nodi fratelli di *d* che vengono ricostruiti a partire dal sotto-albero governato dalla testa *h* di *d*: di nuovo i nodi fratelli vengono divisi in due classi rispetto all'ordinamento lineare della frase.

Lunghezza e direzione di un arco di dipendenza Questo è un tratto complesso che combina due diversi tipi di informazioni: la lunghezza della dipendenza (DL), cioè la distanza lineare tra la testa sintattica *h* e il dipendente *d* (calcolata in termini di parole intercorrenti), e la direzione della dipendenza (DD). Per ogni relazione di dipendenza tra le parole W*i* e W*j*, se W*i* è la testa e W*j* è il dipendente, allora la lunghezza di dipendenza può essere definita come la differenza *i* - *j*. Con questa misura, le parole adiacenti collegate da un link di dipendenza hanno una DL assoluta di 1. Quando *i* è più grande di *j*, la DL è un numero positivo, stando a indicare che la testa si trova dopo il dipendente. Quando *i* è più piccola di *j* la DL corrisponde a un numero negativo.

Mentre a livello delle dipendenze individuali DD è puramente una differenza qualitativa, quando riferita a un corpus analizzato per mezzo di parsing, rappresenta un misura quantitativa utile a misurare alcune caratteristiche sintattiche che caratterizzano alcuni tipi di strutture linguistiche presenti all'interno della collezione di testi.

Plausibilità dell'arco di dipendenza Questo tratto viene utilizzato per calcolare la plausibilità di un arco di dipendenza data la part-of-speech di *d* e di *h,* insieme a quella della testa padre di *h*.

Questo tipo di tratto è stata utilizzato per la prima volta in (Dell'Orletta *e altri*, 2011) per calcolare la bontà del risultato dell'analisi di un parser.

Calcolo della qualità dell'arco sintattico Il punteggio qualitativo di un arco di dipendenza (QS) viene calcolato come una forza associativa che tiene in considerazione il dipendente *d* della sua testa *h* e la testa della testa di *h*.
Rispetto alle *features* sopra descritte la qualità dell'arco viene calcolata all'interno di una funzione che combina i pesi ad esse associati. Questa funzione, descritta in (Dell'Orletta *e altri*, 2013), calcola la probabilità dell'arco attraverso il prodotto dei diversi pesi calcolati rispetto alle caratteristiche linguistiche analizzate in un'ampia quantità di dati automaticamente analizzati da un parser a dipendenze.

4 Analisi dei dati

In questo capitolo verranno fornite le chiavi di lettura del presente progetto di studio attraverso le analisi dei dati estratti tramite le risorse e le procedure linguistico-computazionali illustrate nel capitolo procedente.

4.1 I dati linguistici estratti

Le informazioni morfo-sintattiche e sintattiche dei dati estratti sono state rilevate a partire dall'ordinamento delle relazioni di dipendenza presenti nella treebank *gold* restituito da LISCA. A partire dal modello statistico costruito probabilisticamente sulle features linguistiche osservate nel corpus giornalistico, l'esecuzione dell'algoritmo ha restituito come output la lista di tutte le relazioni di dipedenza distribuite nella IUDT, disposte secondo un ordine descrescente di plausibilità sintattica.

L'output dell'algoritmo è stato organizzato in un documento di testo comprendente, in ordine descrescente, le misure di plausibilità associate da LISCA ad ogni relazione di dipendenza, e relativamente a quest'ultime la loro posizione all'interno della treebank e l'annotazione sintattica e morfosintattica associata sia alla testa che al dipendente.

Le percentuali delle diverse dipendenze e categorie morfo-sintattiche che compongono corpus analizzato è visibile nelle tabelle seguenti.

Dipendenze				POS	
nmod	15,54%	cop	0,99%	Sostantivi	26,41%
case	15,53%	auxpass	0,88%	Preposizioni	16,83%
punct	11,87%	xcomp	0,82%	Verbi	13,28%
det	10,46%	expl	0,77%	Punteggiatura	11,82%
amod	5,92%	nsubjpass	0,76%	Articoli	8,44%
root	4,32%	neg	0,74%	Aggettivi qual.	6,64%
nsubj	4,32%	ccomp	0,55%	Congiunzioni	4,16%
conj	3,71%	mwe	0,37%	Avverbi	4,07%
dobj	3,69%	appos	0,31%	Ausiliari	2,96%
advmod	3,23%	compound	0,30%	Pronomi	2,38%
cc	3,01%	expl	0,25%	Numeri	2,09%
mark	2,35%	iobj	0,15%	Aggettivi det.	2,02%
aux	2,08%	parataxis	0,15%	Clitici	1,73%
nummod	1,57%	csubj	0,13%	Classi residue	0,12%
acl	1,43%	vocative	0,04%	Interiezioni	0,02%
advcl	1,34%	foreign	0,02%		
name	1,18%	discourse	0,02%		

Tabella 4.1 Composizione del corpus di IUDT.

L'output di LISCA, contentente un totale di 246.440 relazioni di dipendenza, è stato suddiviso in 10 fasce di 24,644 relazioni ciascuna (corrispondenti al 10% dell'intero corpus), per poter osservare quali particolari fenomeni linguistici si manifestassero in relazione al decrescere del grado di plausibilità o prototipicità ad essi attribuito, schematizzandone il decorso. Questo ci ha permesso di osservare i fenomeni che occorrono nel passaggio da strutture sintattiche prototipiche a strutture più marcate.

Il *focus* su cui si sono concentrate le nostre analisi è stato il comportamento dell'algoritmo in relazione al trattamento delle parts-of-speech e delle relazioni di dipendenza, di cui sono state considerate le distribuzioni ed alcuni aspetti della struttura sintattica a dipendenze, selezionando e contestualizzando poi un campione specifico di POS e di dipendenze considerate perché particolarmente significative.

Nei prossimi paragrafi verranno riportate le analisi dei dati rispetto a parametri di tipo morfosintattico e sintattico e alla loro distribuzione nelle varie fasce analizzate.

4.2 I tratti morfosintattici

In questo parafrago analizzeremo la distribuzioni delle categorie morfosintattiche dei *dipendenti* all'interno delle relazioni di dipendenza ordinate da LISCA. Nei grafici sono state escluse le parti del discorso interne alle dipendenze di frequenza inferiore all'1%, trattandosi di casi isolati e statisticamente poco rilevanti.

4.2.1 Distribuzione e analisi

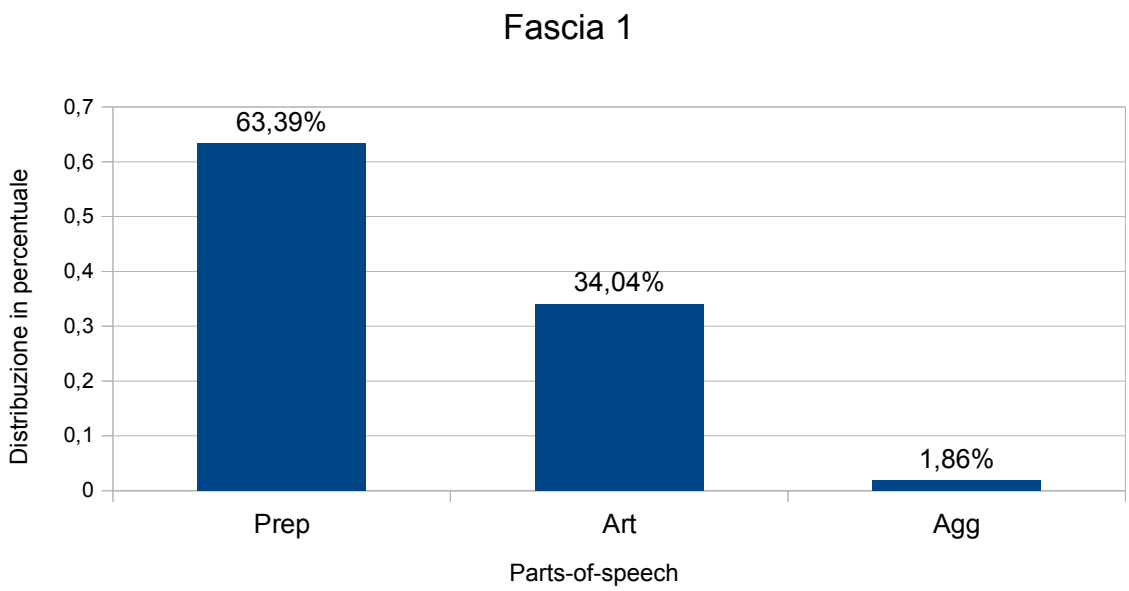
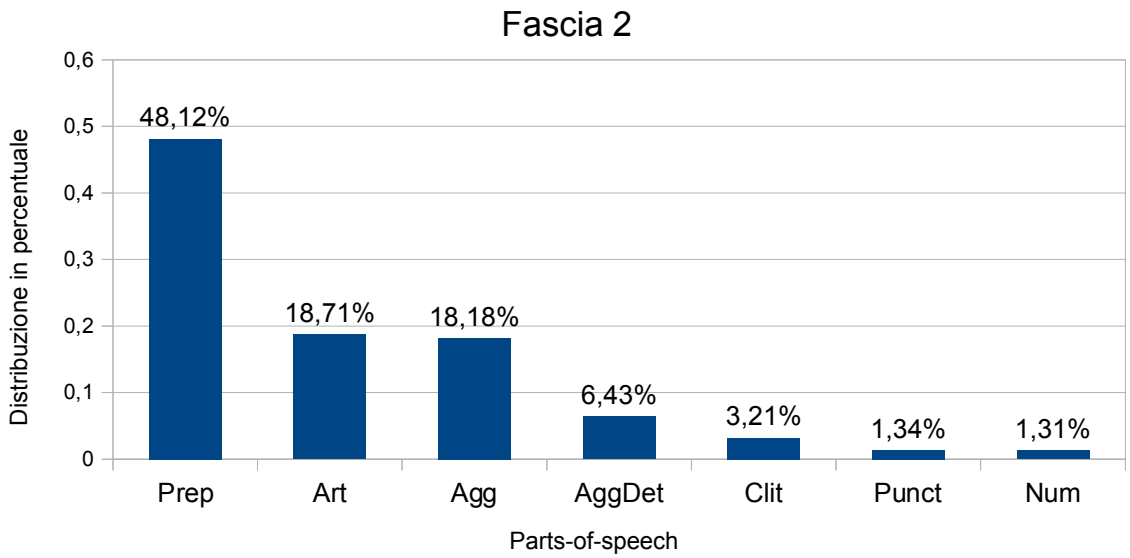
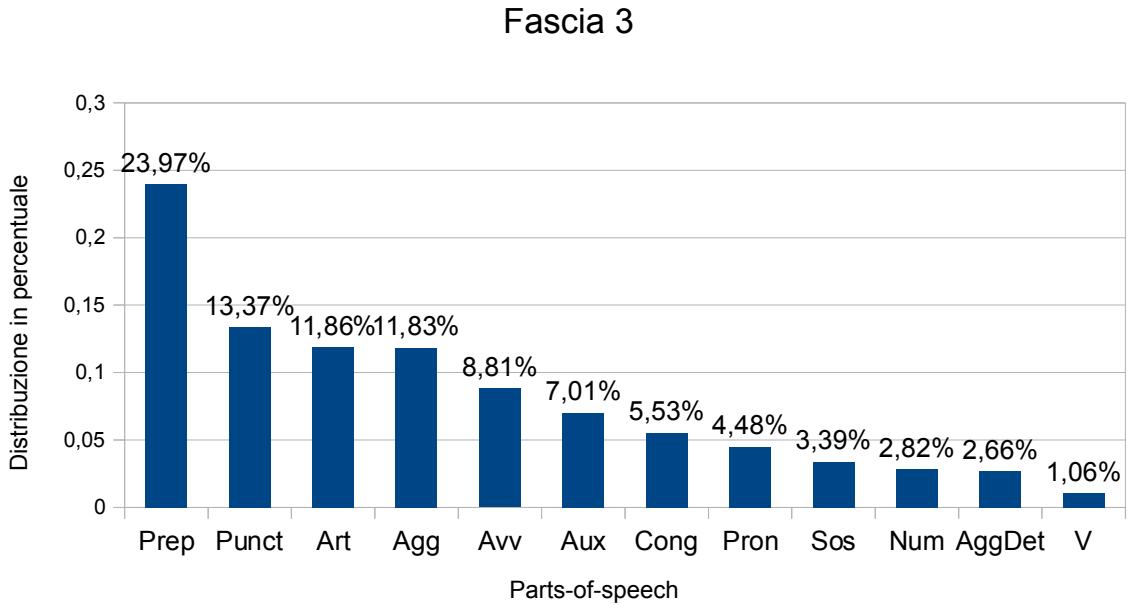
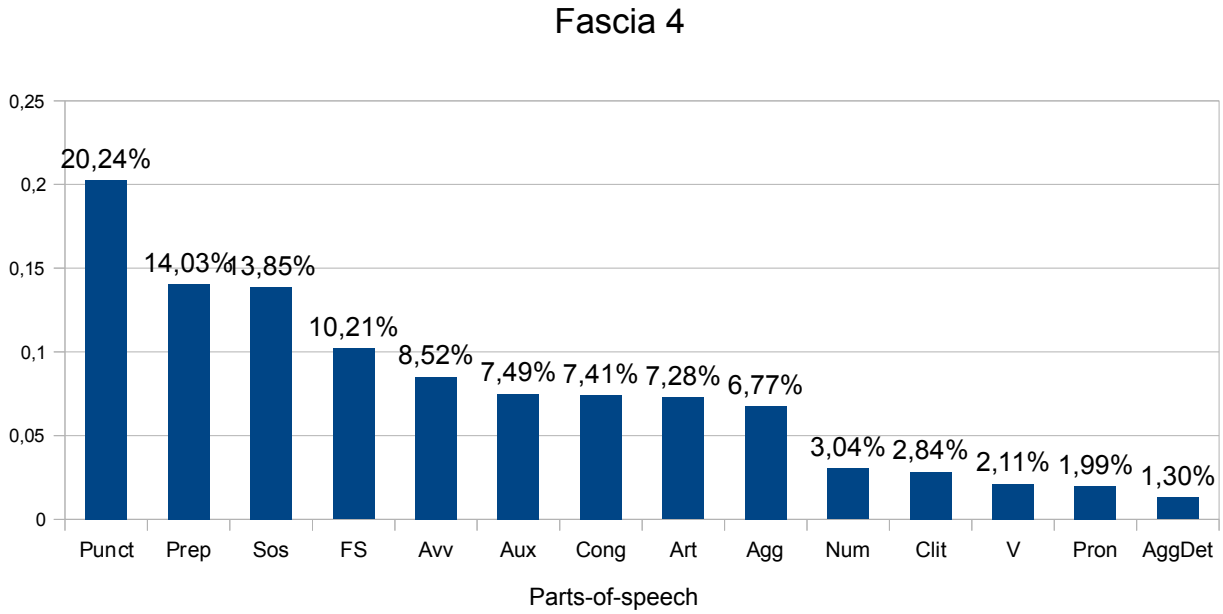


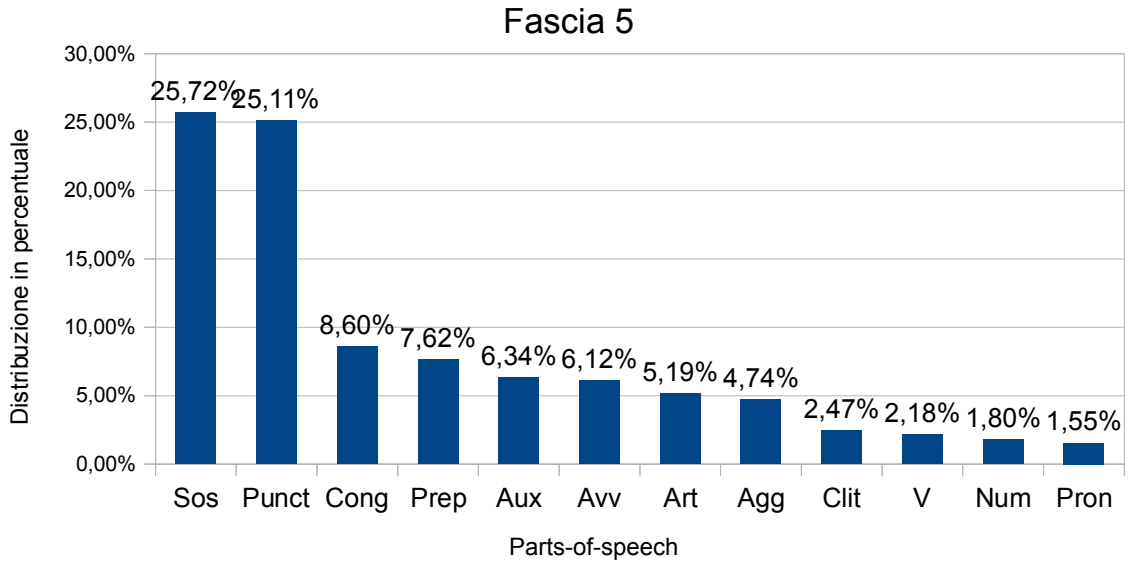
Grafico 4.2.1 Distribuzione delle POS in fascia 1.

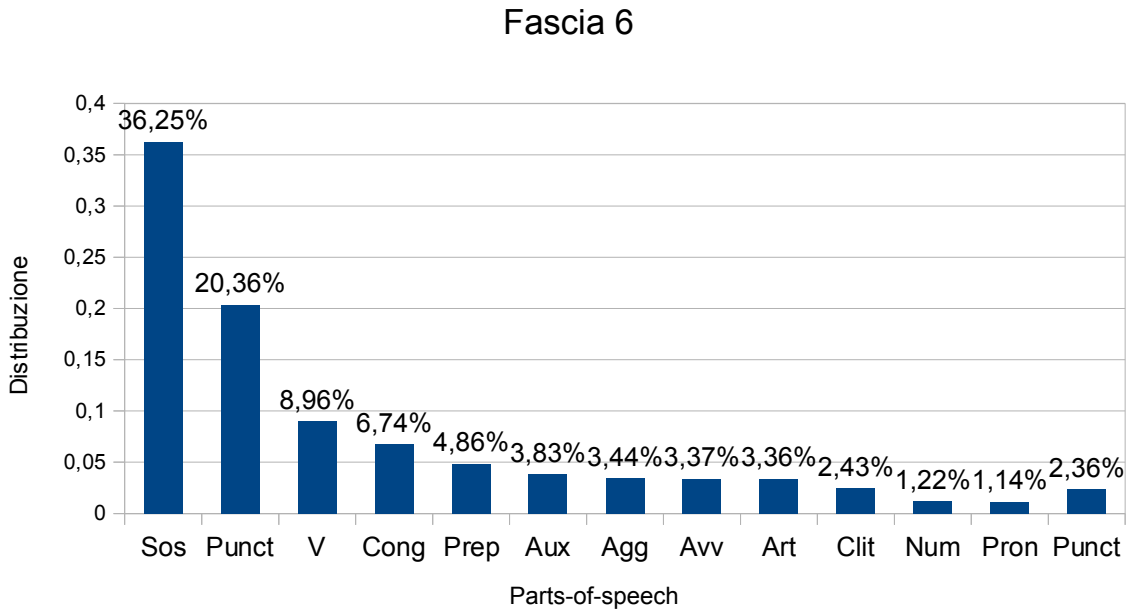


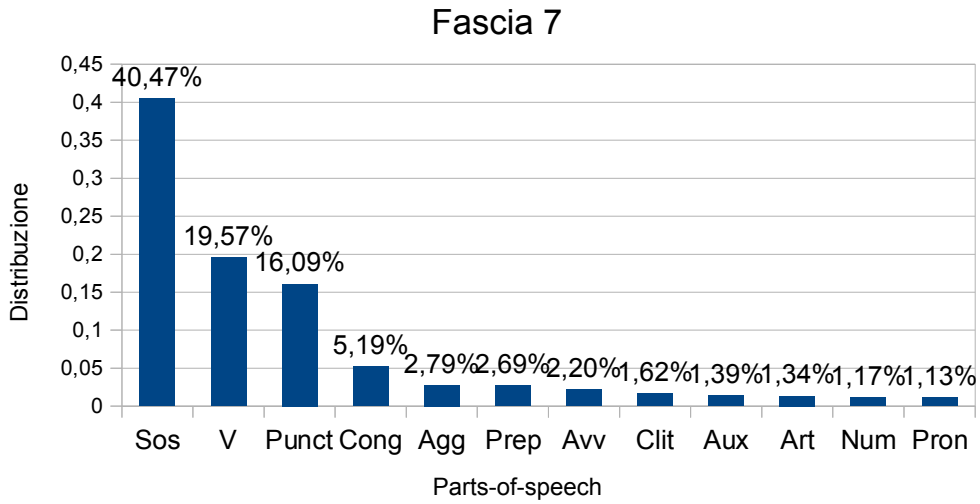




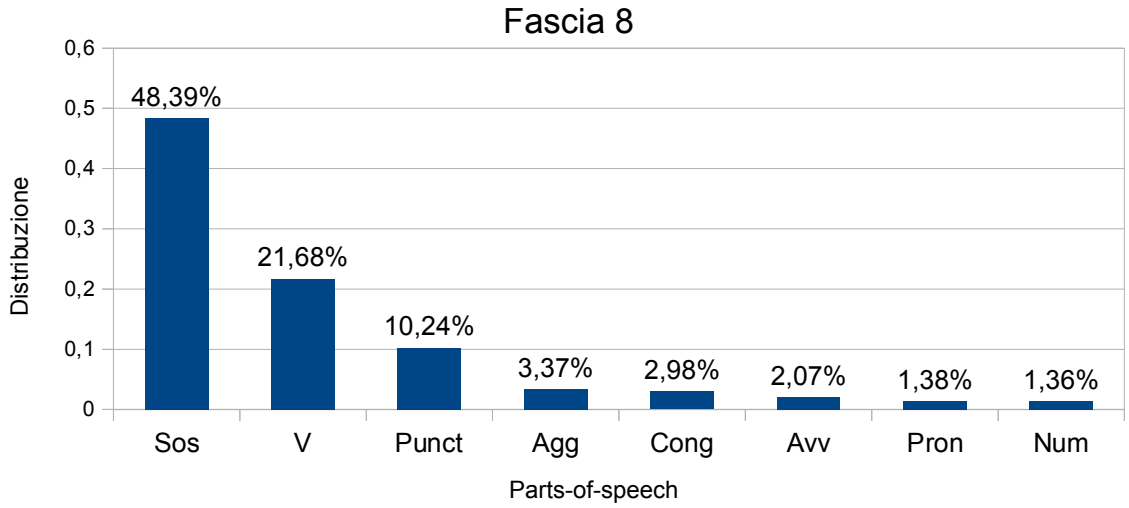
Grafici 4.2.2, 4.2.3, 4.2.4. Distribuzione delle POS in fascia 2, 3 e 4.

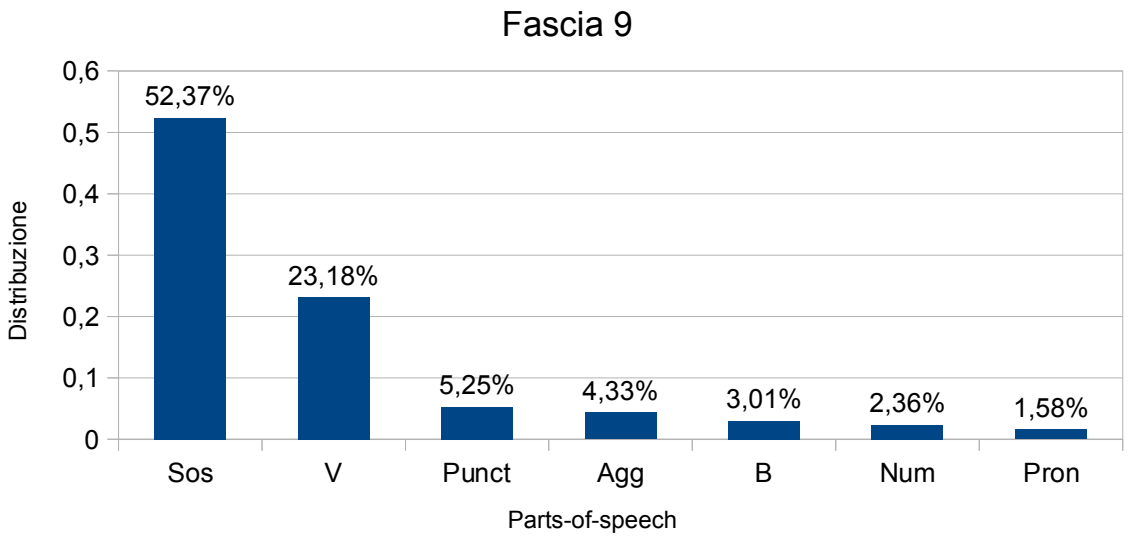


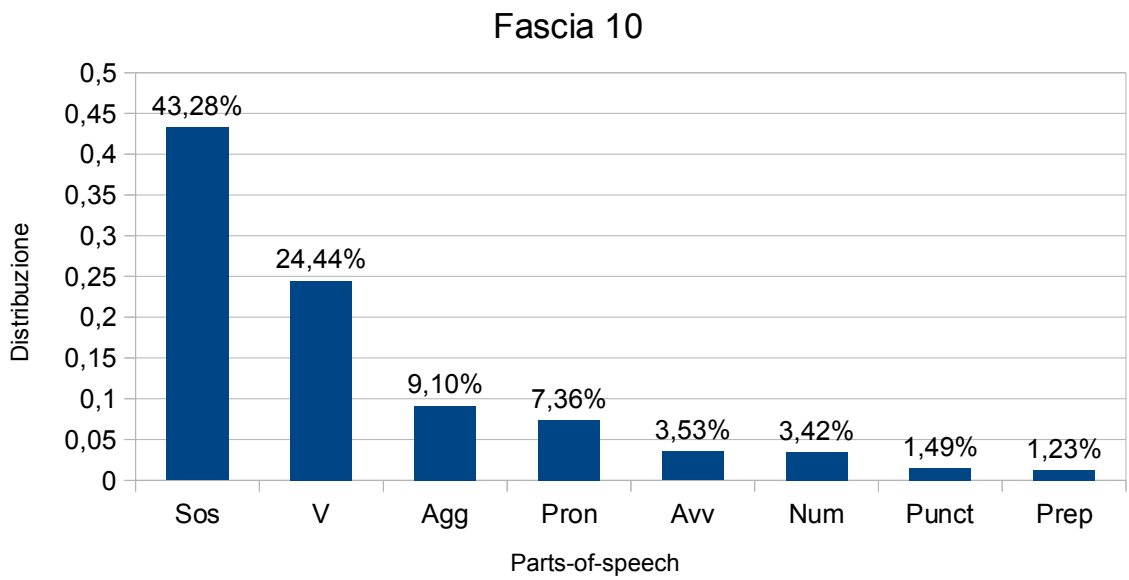




Grafici 4.2.5, 4.2.6, 4.2.7. Distribuzione delle POS in fascia 5, 6 e 7.







Grafici 4.2.8, 4.2.9, 4.2.10. Distribuzione delle POS in fascia 8, 9 e 10.

Come si può osservare dai grafici qui presentati, la distribuzione delle *parti del discorso* attraverso le 10 fasce di analisi risulta disomogenea: il numero di POS individuato tende ad aumentare man mano che il grado di plausibilità degli archi di dipendenza dimiuisce.

Un dato a prima vista osservabile è come le parti del discorso presenti nelle prime fasce siano prevalentemente parole grammaticali (come articoli e preposizioni), numeri o segni di punteggiatura (facili da disambiguare data l'univocità delle funzioni ad esse associate). Andando avanti, nelle fasce intermedie, incominciano a registrarsi, con sempre maggiore incidenza proseguendo verso la fine, parole lessicaoli o semantcamente piene, come sostantivi, pronomi e verbi che compongono per la maggioranza le ultime fasce.

Queste informazioni seggeriscono intuitivamente un determinato dato: le parti del discorso per le quali un sistema di analisi automatico individua le relazioni di dipendenza più facilmente sono in prevalenza elementi sintattici relativamente rigidi, sia in relazione al loro tipo morfologico (le proposizioni sono costituite da morfemi fissi), sia in relazione alla posizione che assumono rispetto alla loro testa (sia gli articoli, sia le preposizioni si trovano quasi sempre in posizione antecedente alla parola-testa di cui sono dipendenti). Al contrario, gli elementi considerati più complessi da analizzare, dunque diffilmente riconosciuti da LISCA come corretti, sono le parti del discorso più flessibili, a livello sia morfologico che sintattico. Verbi e sostantivi, in quanto parole semanticamente piene rendono più ambigua la loro interpretazione ed annotazione. Un'ulteriore conferma di questa suddivisione è data dalla distribuzione degll'aggettivo, presente in maniera più o meno omogenea nelle varie le fasce: nonostante sia una delle tre POS osservate nella prima fascia, la sua occorrenza attraverso le fasce rimane in percentuale costante, rivelando una complessità di riconoscimento intermedia. Gli aggettivi infatti, sebbene svolgano una funzione fissa (modificare semanticamente un'altra parte del discorso), che li porta a disporsi vicino alla testa che modificano e di cui sono dipendenti, possono, nella loro analisi, risultare come ambigui, soprattutto quando sono interni a strutture di coordinazione o si trovano distanti rispetto alla loro testa. Sono questi i casi che portano alcuni aggettivi a collocarsi nelle ultime fasce.

Quanto detto fino ad ora viene confermato dal grafico 4.2.11 dove si può osservare l'andamento distribuzionale di quattro parti del discorso: articoli (deteminativi e non determinativi), preposizioni (articolate e non articolate), sostantivi (nomi propri e comuni) e aggettivi. Mentre le prime due si concentrano nelle fasce conontenenti le relazioni più plausibili, e le loro occorrenze vanno diminuendo fino a raggiungere numeri di occorrenza irrilevanti, facilmente riconducibili al margine di errore previsto dall'utilizzo di metodi di analisi statistica, gli aggettivi, distribuiti omogeneamente, rivelano una complessità intermedia, in quanto tra le relazioni che li coinvolgono si distinguono in egual numero casi di maggiore o minore plausibilità sintattica. I sostantivi, invece,

risultano inversamente proporzionali alla distribuzione delle parole grammaticali, collocandosi tra le parti del discorso più ambigue e complesse da costestualizzare all'interno della frase: sono del tutto assenti nelle prime fasce, aumentano gradualmente a partire dalla terza fascia e costituisco quasi il 45% dell'ultima.

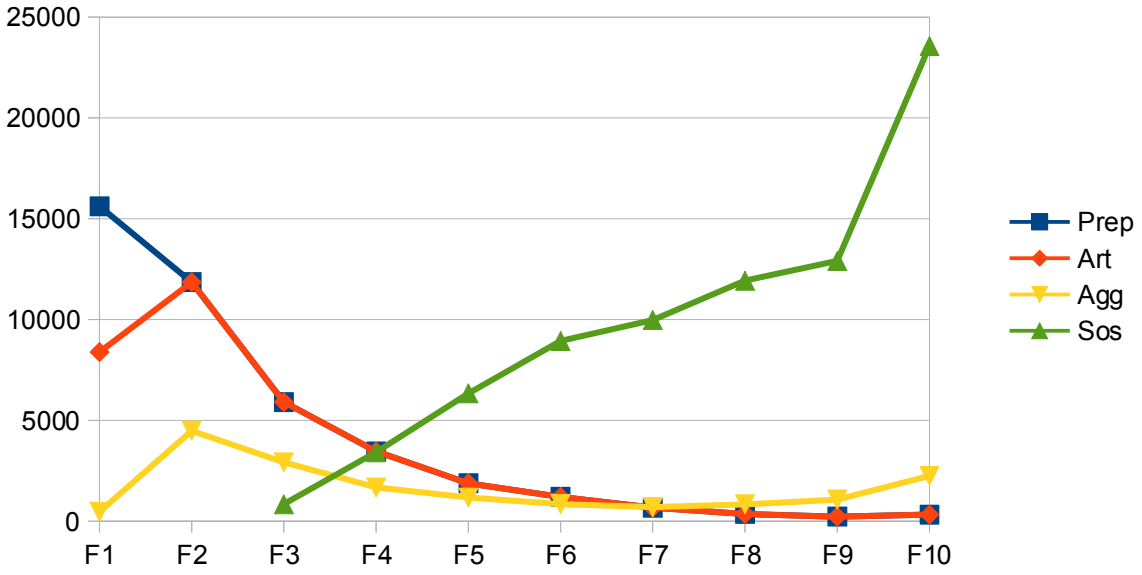


Grafico 4.2.11 Distribuzione attraverso le fasce di quattro parti del discorso: preposizioni, articolo, aggettivi e sostantivi

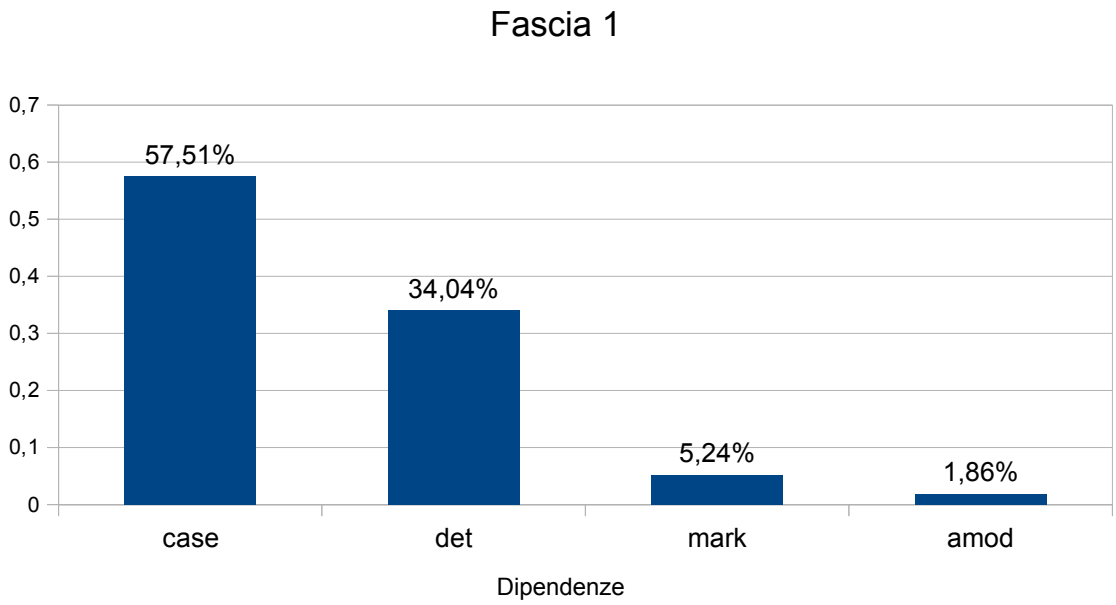
In termini di complessità computazionale il risultato può essere riassunto nel modo seguente: a livello automatico le parti del discorso delle costruzioni sintattiche più facilmente riconoscibili come corrette, sono le parole grammaticali (o *function words,* parole semanticamente vuote e generalmente meno marcate morfologicamente), le quali possono essere considerate meno complesse in quanto funzionali allo svolgimento di un ruolo sintattico univoco e tendenzialmente disposte secondo un ordine fisso che le colloca immediatamente adiacenti o di poco distanti alle parole piene da cui dipendono e che ne completano il significato. Le parole piene invece, proprio perché semanticamente più indipendenti e morfologicamente più complesse, risultano più ambigue e difficilmente contestualizzabili all'interno delle strutture sintattiche che le ricomprendono.

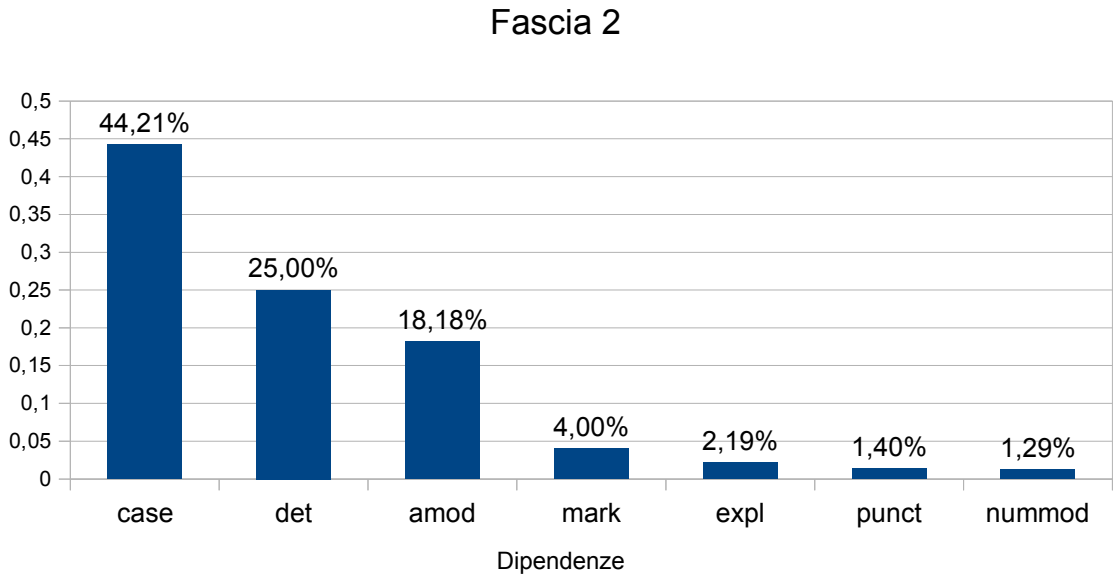
4.3 I tratti sintattici

In questa terza parte del capitolo verranno analizzate le relazioni di dipendenza ordinate da LISCA,

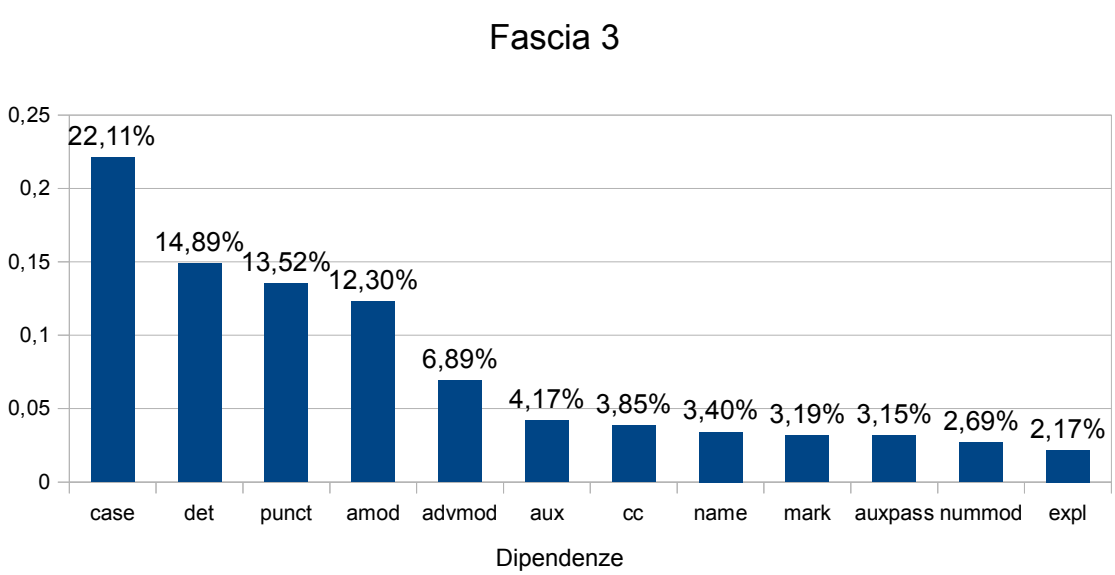
osservandone le distribuzioni ed alcuni tratti sintattici. Nella rappresentazione delle distribuzioni sono state escluse le dipendenze la cui frequenza fosse inferiore all'1-2%, trattandosi di casi isolati e statisticamente poco rilevanti.

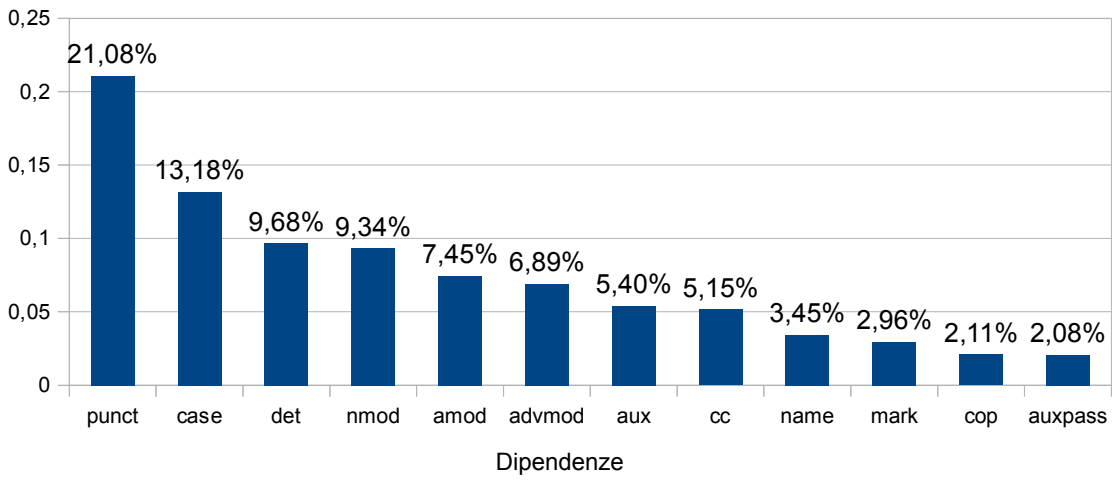
4.3.1 Distribuzione e analisi

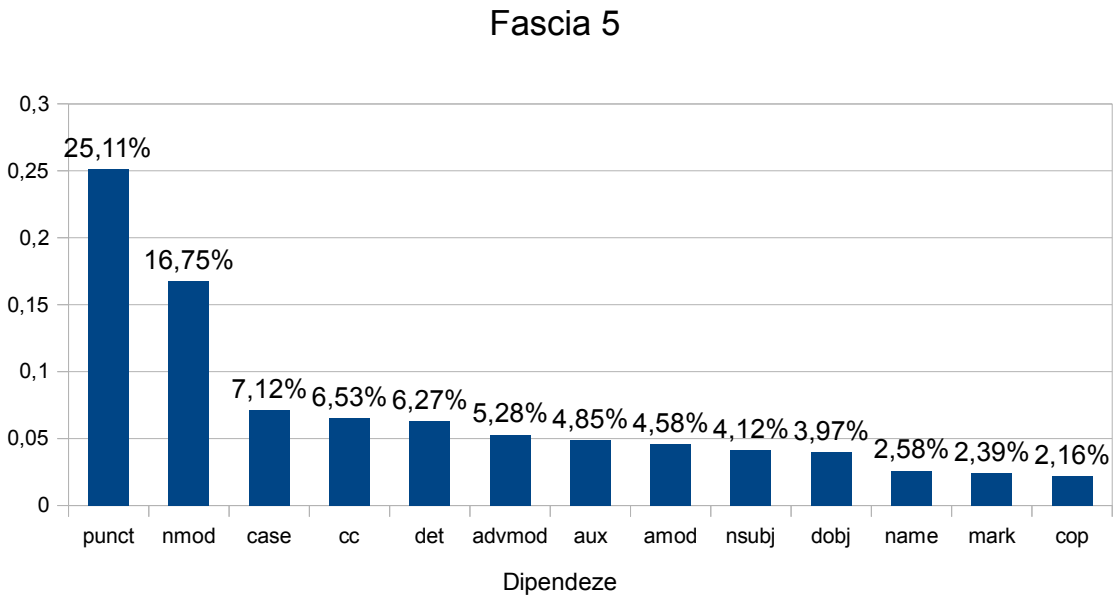




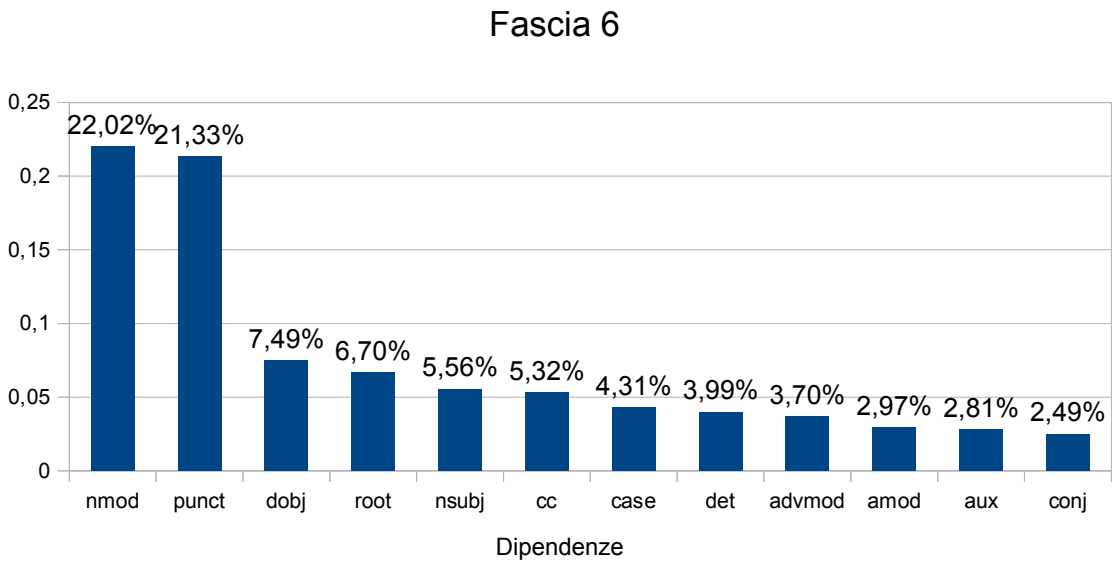
Grafici 4.3.1, 4.3.2 Distribuzione delle relazioni di dipendenza in fascia 1 e 2.

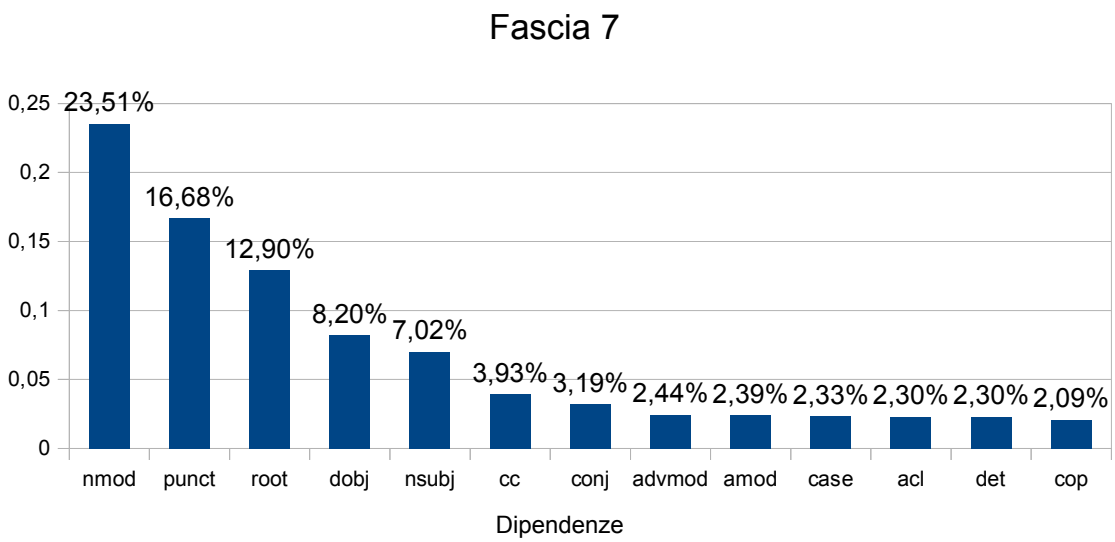


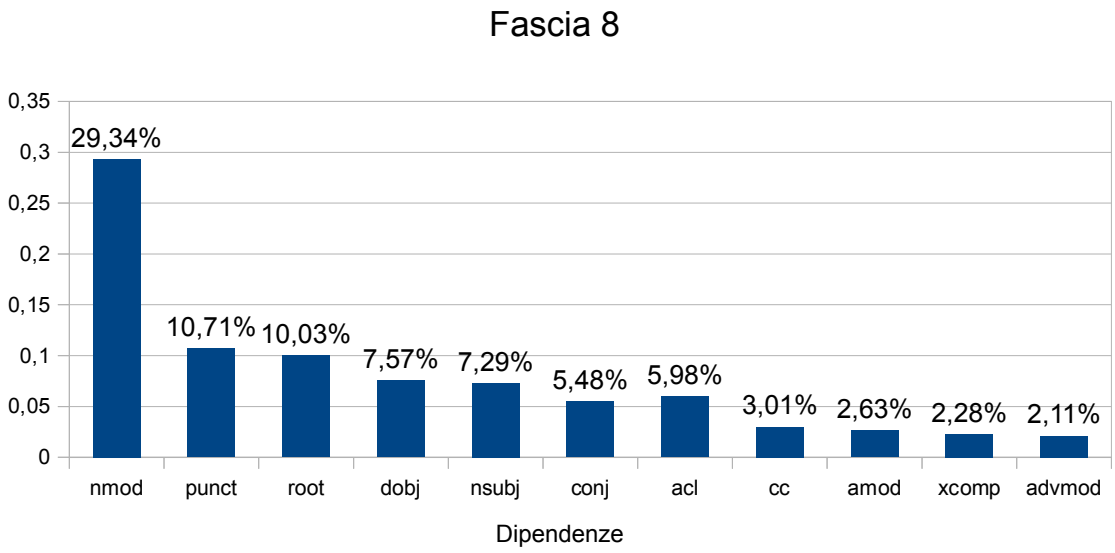




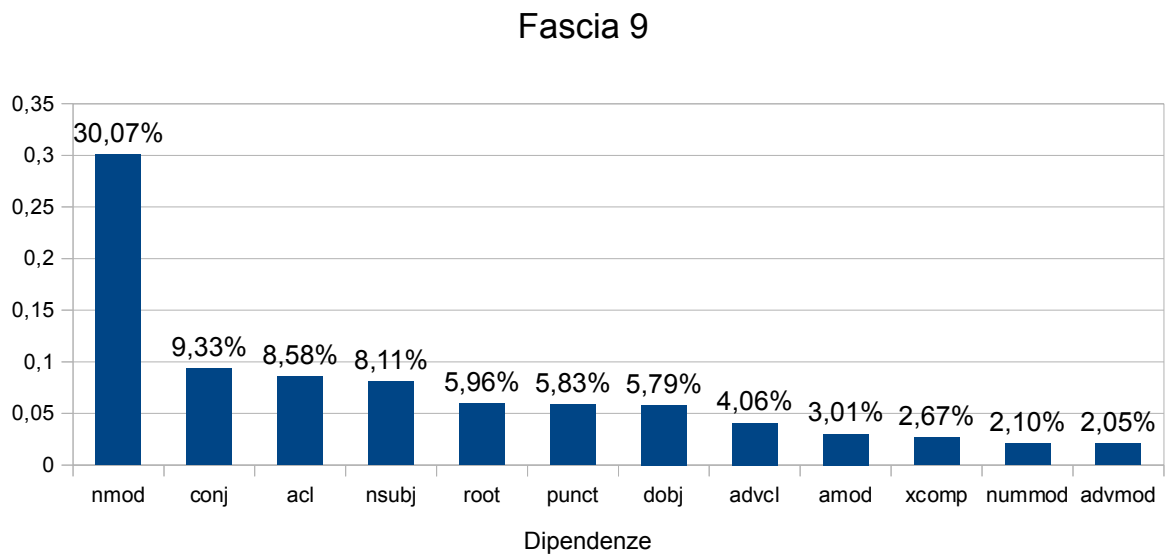
Grafici 4.3.3, 4.3.4, 4.3.5 Distribuzione delle relazioni di dipendenza in fascia 3, 4 e 5.

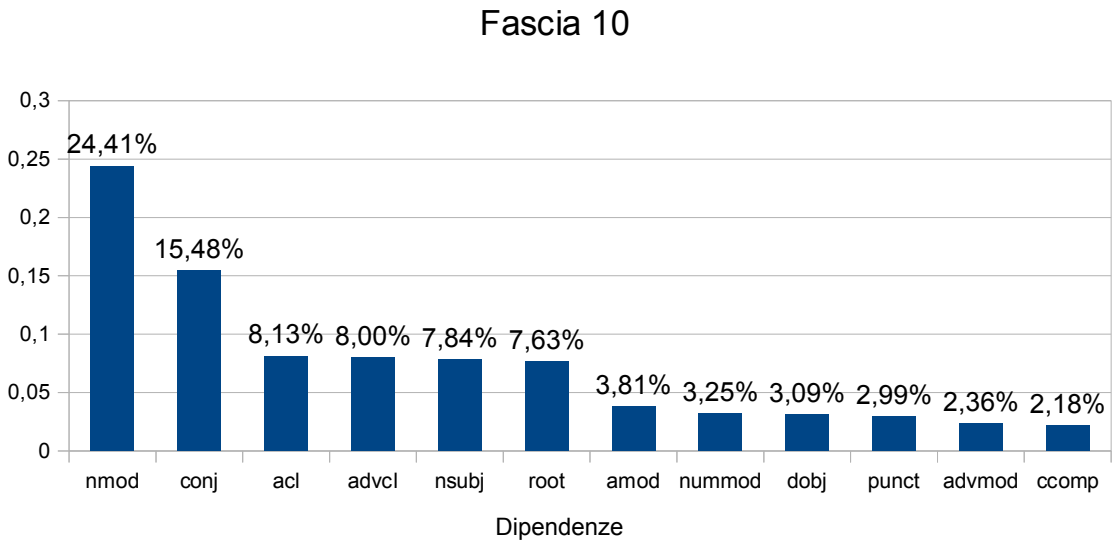






Grafici 4.3.6, 4.3.7, 4.3.8 Distribuzione delle relazioni di dipendenza in fascia 6, 7 e 8.





Grafici 4.3.9, 4.3.10 Distribuzione delle relazioni di dipendenza in fascia 9 e 10.

In questi grafici, rappresentanti la distribuzione delle singole relazioni di dipendenza per ogni fascia, si osserva un andamento simile a quello registrato per le categorie morfosintattiche: nelle prime fasce tende a concentrarsi un nucleo fisso di dipendenze, le cui occorrenze diminuiscono in maniera direttamente proporzionale al grado di plausibilità degli archi considerati e parallelamente al registrarsi di nuove dipendenze, sintatticamente più complesse.

Come era intuibile dalle distribuzioni delle POS, gli archi riconosciuti come più prototipici sono quelli che realizzano relazioni tra parole grammaticali e parole piene, come *det*, *case* e *mark*, che collegano rispettivamente articoli, preposizioni e congiunzioni subordinanti o avverbi alla loro testa sintattica. Nelle prime fasce la testa è rappresentata nella maggior parte dei casi da sostantivi,

numeri e aggettivi nelle relazioni di tipo *det*, da sostantivi, verbi, aggettivi e pronomi nelle dipendenze di tipo *case*.

Proseguendo oltre le prime fasce le relazioni che vengono osservate oltre quelle sopracitate, sono descritte da archi che, tranne nel caso delle congiunzioni (*cc*), non collegano parole grammaticali, ma individuano particolari tipi di parole lessicali. Si tratta di *advmod*, *nummod*, *cop*, e *aux*, che collegano rispettivamente avverbi, numeri (cardinali e ordinali), copule e ausiliari (verbi modali compresi) alla loro testa. La presenza di queste relazioni nelle prime fasce si può spiegare in termini di marcatezza linguistica: per quanto riguarda avverbi e numeri, la loro morfologia, generalmente più rigida e meno marcata, li rende facilmente disambiguabili; lo stesso vale per copule e ausiliari che, oltre ad essere categorie verbali *ristrette,* cioè ricollegabili a un numero finito di verbi, presentano un ordine non marcato e per questo sono facilmente riconoscibili dal contesto (le copule sono sempre adiacenti alla loro testa o, in assenza di quest'ultima, al loro dipendente; gli ausiliari invece sono sempre antecedenti al verbo che li completano semanticamente).

Un discorso a parte va fatto per le relazioni di *amod* e *nmod* che risultano distribuite omogeneamente in tutte le fasce. Queste due relazioni, utilizate per collegare modificatori aggettivali o nominali alle parole-testa che modficano, presentano una complessità intermedia: come abbiamo visto nel paragrafo precedente relativamente agli aggettivi, le relazioni che coinvolgono i modificatori del nome rappresentano fenomeni linguistici la cui struttura, a seconda dei contesti, può risultare altamente semplice e prototipica o come altamente complessa e poco plausibile. Questo tratto è ricollegabile al fatto che le unità linguistiche coinvolte in questo tipo di dipendenze possono disporsi secondo un ordine più libero e dunque possibilmente più marcato rispetto ad altre parti del discorso. *Nmod* presenta una complessità maggiore rispetto ad *amod* perché conivolge modificatori nominali, tendenzialmente più liberi e dunque più ambigui di quelli aggettivali.

Quanto detto fino ad ora vale anche e più di tutto per le costruzioni sintattiche che troviamo solamente a partire dalla quinta fascia e che si concentrano maggiormente verso la fine: *nsubj*, *dobj*, *ccomp*, *xcomp* e *root*. Queste relazioni, riconosciute da LISCA come più complesse e meno plausibili, sono non a caso le relazioni che collegano parole lessicali piene come sostantivi e verbi alla loro testa. I livelli di complessità individuati da questo tipo di parole sono diversi: oltre al già citato aspetto morfologico, le parole lessicali tendono ad inserirsi in sistemi gerarchici più complessi in quanto possono fare contemporatemente da testa e da dipendente; inoltre hanno maggior libertà di movimento all'interno della frase, nel senso che non hanno un ordine fisso e la loro interpretazione è più ambigua.

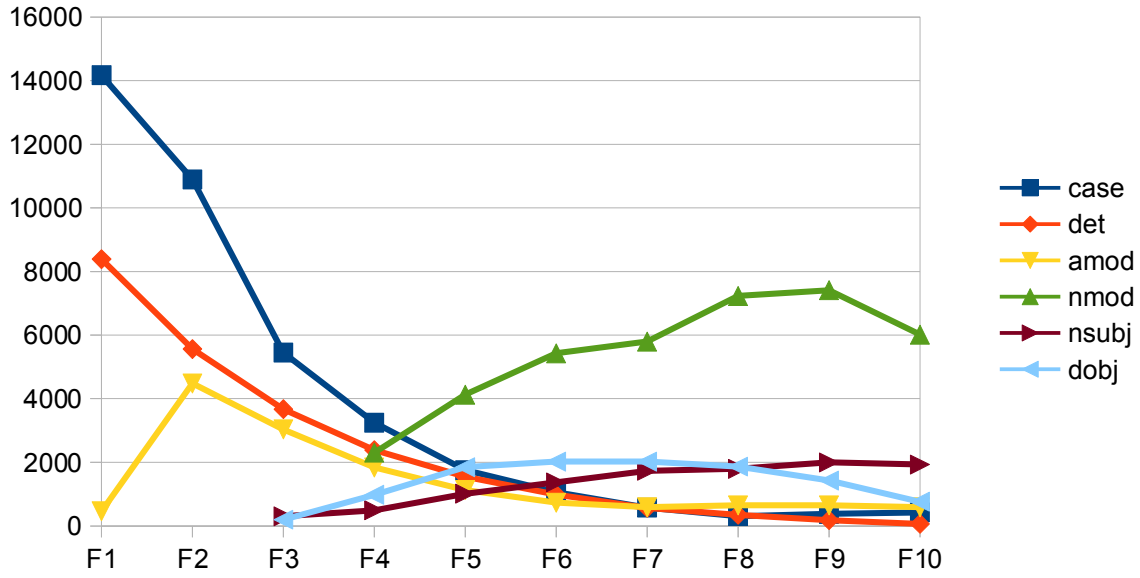


Grafico 4.3.11 Distribuzioni di sei relazioni di dipendenza: *case, det, amod, nmod, nsubj e dobj.*

Nel grafico 4.3.11 sono state raccolte le distribuzioni di sei relazioni di dipendenza per osservarne meglio l'andamento: sono state selezionate due dipendenze definibili come semplici (det, case), due esempi di relazioni con complessità *intermedia* (amod, nmod) e due relazioni complesse (nsubj, dobj).

4.3.2 Orientamento delle strutture sintattiche

Un altro tratto sintattico che abbiamo utilizzato come parametro di analisi è stato l'orientamento delle strutture sintattiche, osservando la direzione degli archi dell'albero sintattico a dipendenze. Questa direzione viene definita in base ai due possibili orientamenti (destra/sinistra) che le relazioni di dipendeza possono assumere nell'ordine lineare della frase: a seconda che il dipendente di una particolare costruzione sintattica si trovi a destra o a sinistra della testa, la direzione dell'arco che lo collega sarà orientata verso la sua testa.

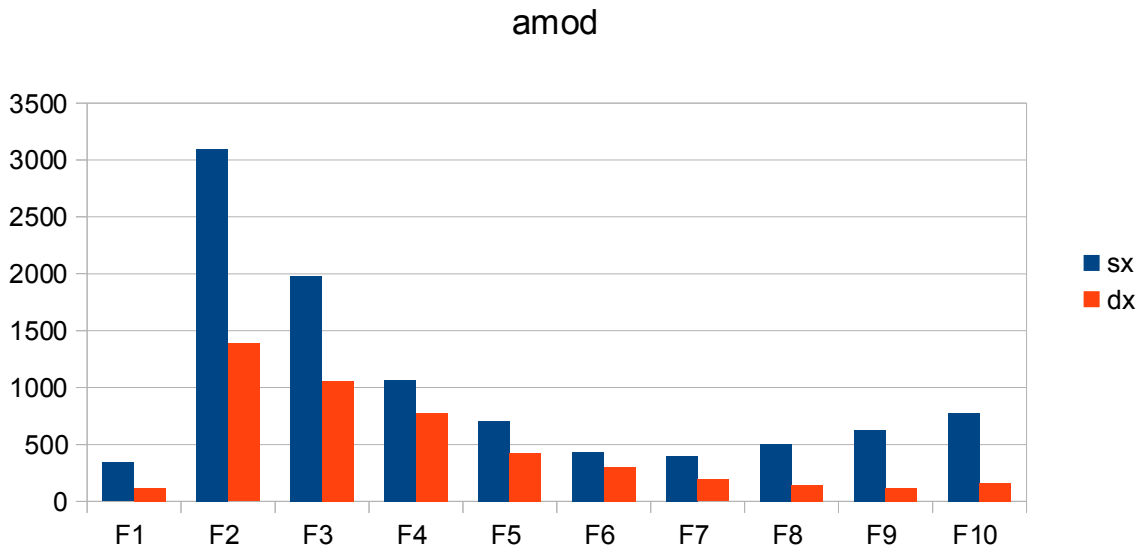
Da un punto di vista tipologico si tratta di un parametro particolarmente rilevante in termini di complessità, soprattutto per quanto riguarda l'ordine dei costituenti: questo studio ci ha permesso di osservare come gli elementi più marcati sintatticamente, cioè dislocati rispetto al loro ordine canonico, costituiscono anche le parti del discorso più complesse per l'algoritmo di LISCA. Al contrario le relazioni che vengono valutate come più plausibili e prototipiche sono quelle che collegano le parti del discorso meno marcate.

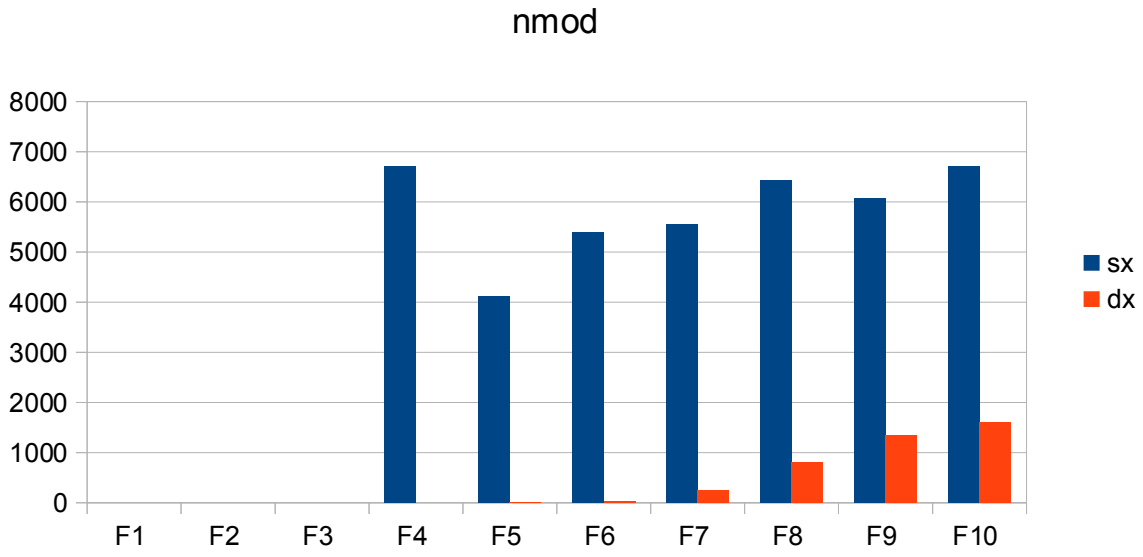
Questo è il caso di alcuni modificatori del nome, in cui la deviazione rispetto alla testa non permette

eccezioni e non può essere soggetta a condizionamenti di natura pragmatica: gli articoli, i determinanti, i quantificatori e i numerali risultano più plausibili perché si trovano sempre in una posizione antecedente alla loro testa.

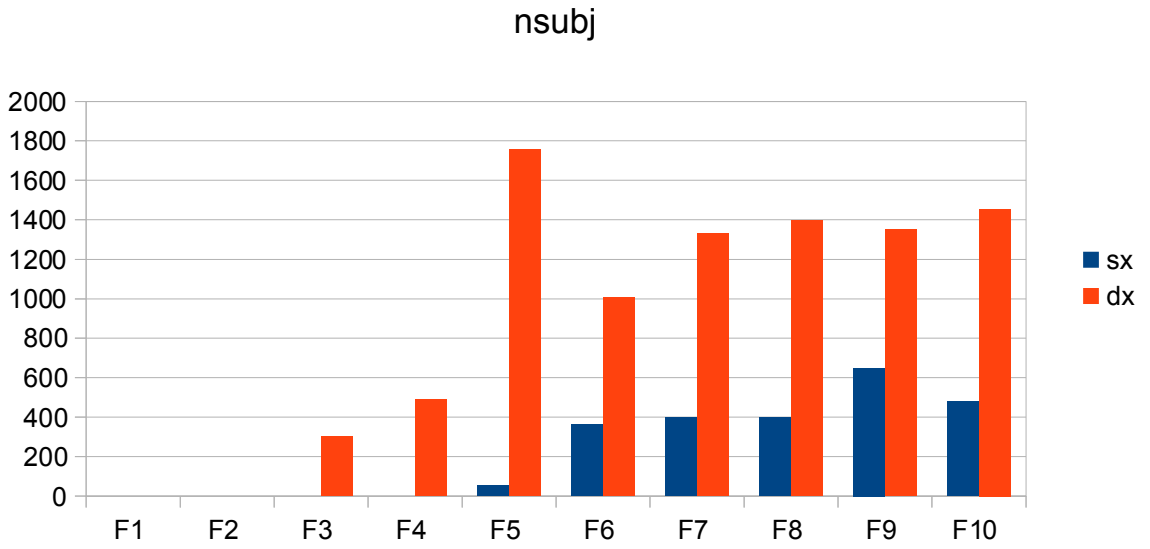
E' interessante notare come nell'ordinamento prodotto da LISCA le relazioni *det* e *case*, che collegano articoli e preposizioni alla loro testa, e che costituiscono per l'algoritmo le dipendenze più prototipiche e semplici, indicano la testa sempre a destra, senza eccezioni.

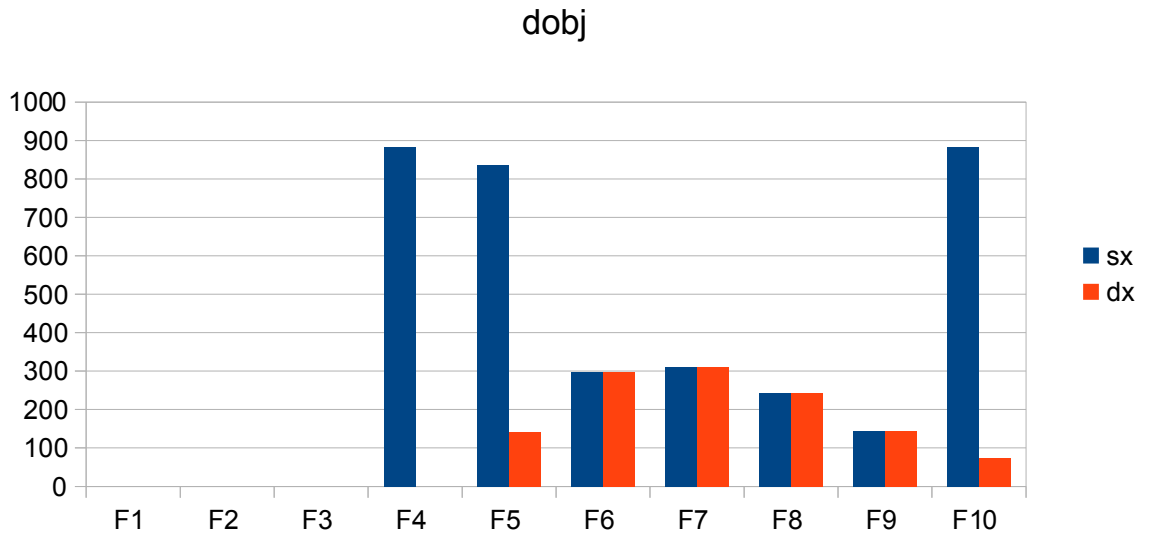
Nei grafici seguenti (4.3.12, 4.3.13, 4.3.14, 4.3.15) sono state rappresentate le distribuzioni di quattro particolari relazioni sintattiche: *amod*, *nmod*, *nsubj* e *dobj,* discriminate in base alla direzione della testa.





Grafici 4.3.12, 4.3.13 Distribuzioni di *amod* e *nmod,* in funzione della direzione della testa.





Grafici 4.3.14, 4.3.15 Distribuzioni di *nsubj* e *dobj,* in funzione della direzione della testa.

In tutti i casi la maggior parte dei dati rilevati rispecchia a livello tipologico l'ordine canonico dei costituenti. Nel caso di *amod*, che avevamo detto distribuirsi omogeneamente attraverso le fasce, le teste dei modificatori aggettivali si trovano in netta maggioranza a sinistra, soprattutto nella seconda fascia dove il grado di prototipicità è più alto. Questo è indice del fatto che le costruzioni sintattiche che l'algoritmo riconosce come più plausibili sono quelle meno marcate in quanto rispettano l'ordine canonico: infatti l'aggettivo non ha un'ordine fisso, ma la sua posizione *non marcata* viene fatta coincidere con quella dopo il nome cui si riferisce perché, quando un aggettivo qualificativo precede il nome, esso indica di solito una maggiore soggettività di giudizio in chi parla o scrive, una

particolare enfasi o ricercatezza stilistica (Treccani, 2011).

La relazione distribuita in maniera più netta rispetto alla sua direzione è quella di *nmod,* che collega alla loro testa i modificatori nominali: in quasi tutti i casi la testa si trova a sinistra, secondo l'ordine canonico tema-rema.

Anche le distribuzioni di *nsubj* e *dobj* rispecchiano in modo abbastanza diretto l'ordine di tipo Soggetto-Verbo-Oggetto: *nsubj*, che collega il soggetto al verbo da cui dipende, nella maggior parte dei casi presenta la testa a destra, confermando la tendenza dei soggetti ad occupare una posizione preverbale; invece nelle relazioni di tipo *dobj,* che generalmente collegano l'oggetto diretto al verbo, nonostante non prevalda in maniera netta un determinato orientamento sull'altro (infatti il complemento oggetto in italiano si può trovare in posizione preverbale o postverbale), le costruzioni presentano più frequentemente la testa a sinistra, come nell'ordine canonico SVO.

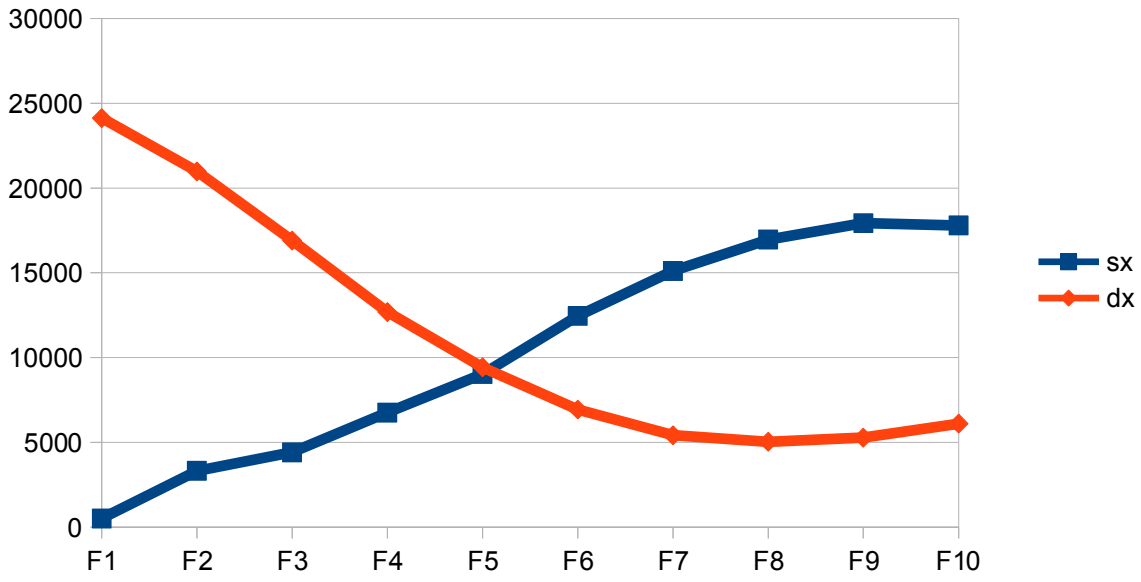


Grafico 4.3.16 Distribuzioni delle relazioni di dipendenza, in funzione della direzione della testa.

Infine, nel grafico 4.3.16, in cui sono state rappresentate le distribuzioni di tutte le relazioni di dipendenza (eslusa la punteggiatura), facendo distizione tra dipendenze con testa a destra e dipendenze con testa a sinistra, si può osservare come i due tipi di orientamenti (destra/sinitra), nonostante occorrano con una frequenza simile, vengano descritti da andamenti opposti: nelle prime fasce si concentrano le costruzioni con testa a destra, nelle ultime quelle con testa a sinistra.

Si nota come le costruzioni più complesse siano quelle con testa a sinistra, mentre quelle più semplici, che si concentrano nelle prime fasce, hanno testa a destra: questo dato è da ricordarsi al fatto che la maggior parte delle costruzioni che presentano testa a sinistra (sintagmi verbali,

sintagmi nominali del tipo “nome + aggettivo”, “nome + frase relativa” ecc.) sono anche le più difficili da disambiguare, mentre le costruzioni con testa sempre a destra (come quelle che legano i determinanti alle loro teste) presentano un ordinamento fisso e dunque computativamente più prevedibile.

4.3.3 Lunghezza delle relazioni di dipendeze

L'ultimo parametro utilizzato per una valutazione di complessità sintattica è stata la lunghezza dei link di dipendeza (calcolata come la distanza (in parole) tra la testa e il dipendente). Infatti un altro fattore di complessità ampiamente riconosciuto nella letteratura linguistica, psicolinguistica e linguistico-computazionale (cfr. Lin, 1996; Gibson, 1998) riguarda la “misura” della lunghezza delle relazioni di dipendenza. Per poter analizzare questo dato esclusivamente in relazione alle dipendenze tra parole, da tutte le rappresentazioni di questo pararagrafo sono state escluse le dipendenze di punteggiatura.

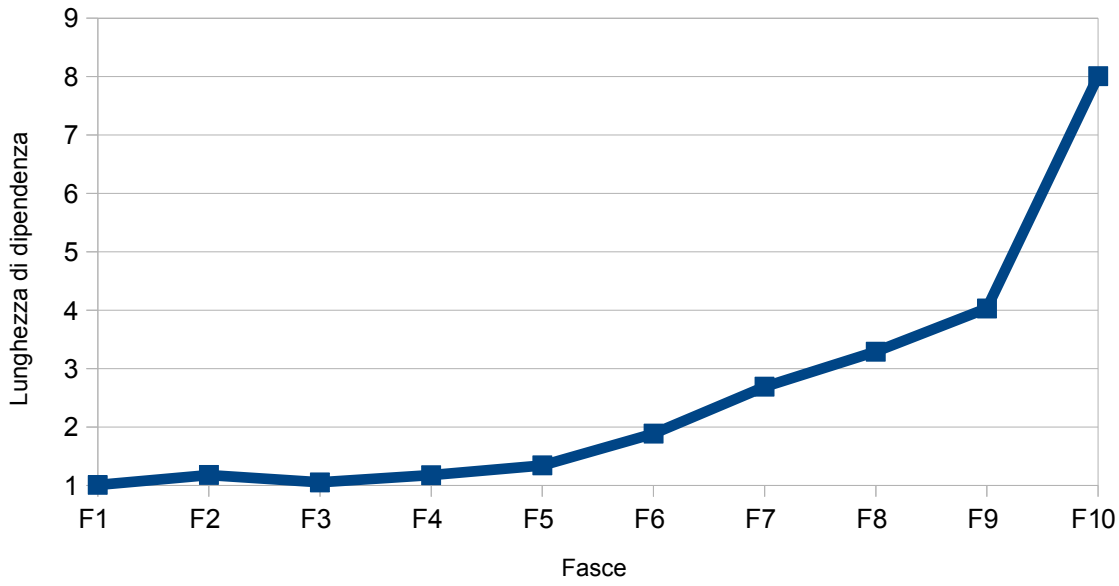


Grafico 4.3.17 Media delle lunghezze di dipendenza attraverso le fasce.

Calcolando la media delle lunghezze dei link di dipendenza per ogni fascia, mostrate nel grafico x, si osserva chiaramente come il grado di complessità attribuito da LISCA agli archi di dipendenza cresca in parallelo alla media delle distanze: le relazioni in cui la distanza tra testa e dipendente è minore sono quelle più prototipiche, mentre le dipendenze mediamente più lunghe si distribuiscono maggiormente nelle fasce che raccolgono le relazioni più complesse.

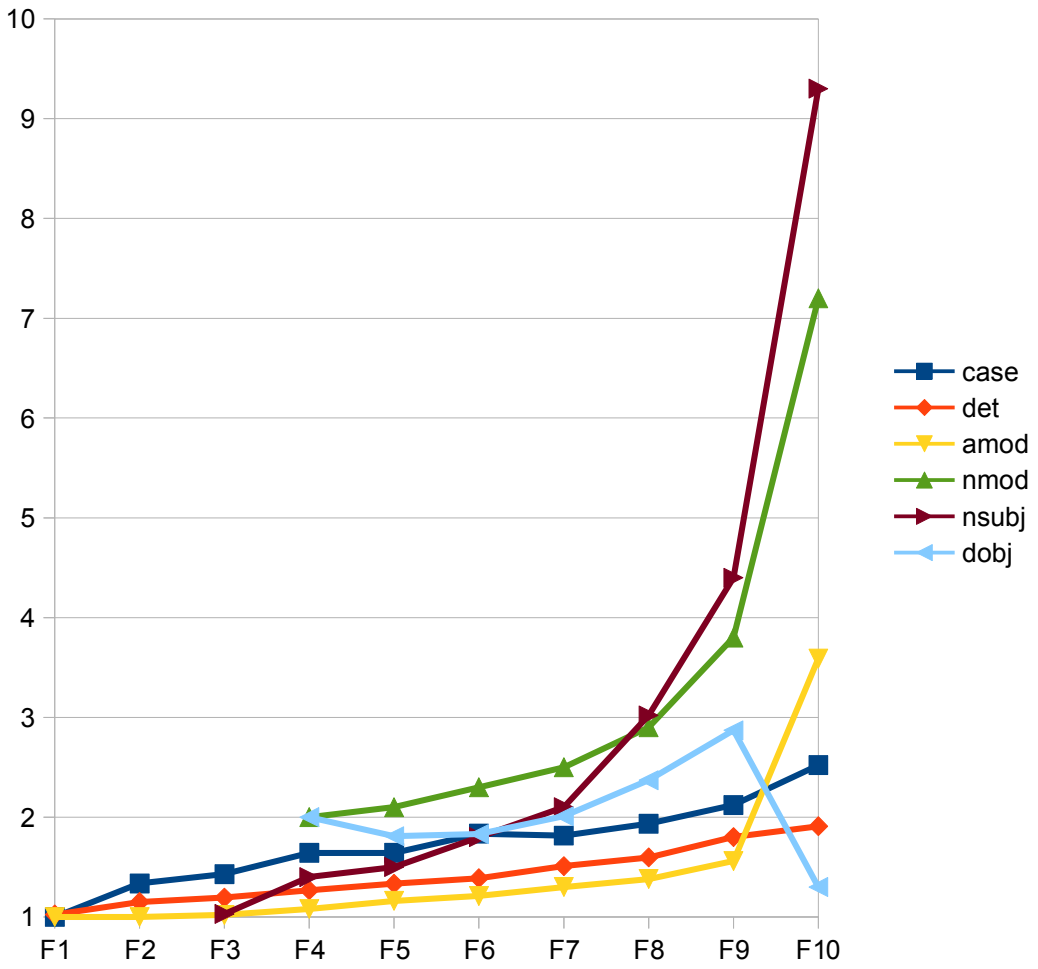


Grafico 4.3.18 Medie delle lunghezze di sei relazioni di dipendenza: *det, case, amod, nmod, dobj* e *nsubj.*

Nel grafico 4.3.18 sono state riportate le medie delle lunghezze delle relazioni sintattiche su cui fino ad ora ci siamo concentrati, apportando un'ulteriore conferma a quanto detto fino ad ora: le dipendenze più prototipiche, *det* e *case*, ma anche quelle meno prototipiche come *amod* e *dobj*, che intercorrono tra determinati o modificatori o oggetti diretti, e che a livello semantico sono strettamente legate alla testa da cui dipendono, sono anche 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, come abbiamo visto, collegano sintagmi più flessibili, cioè con una maggiore libertà di movimento rispetto ad altre parti del discorso, raggiungono in media, soprattutto verso le ultime fasce, le lunghezze maggiori, coerentemente al grafico visto precedentemente.

Conclusione

In questo studio è stata proposta una nuova metodologia d'analisi per lo studio computazionale della complessità sintattica: per poter analizzare la complessità abbiamo osservato il passaggio dalla prototipicità per un sistema computazionale automatico alla nozione di marcatezza linguistica. Le analisi sono state rese possibili tramite l'applicazione di una algoritmo di analisi linguistica, LISCA, utilizzato per misurare il grado di plausibilità e di prototipicità delle strutture sintattiche all'interno della IUDT, treebank *gold* dell'italiano.

L'elaborazione dei dati necessari alla realizzazione di questo studio si è servita di una precisa metodologia d'analisi in cui si distinguono tre fondamentali passaggi:

- **Passo 1**: annotazione linguistica automatica di un grande corpus di testi giornalistici italiani in una procedura a tre fasi (*sentence splitting*, *POS tagging* e *syntactic parsing*). Gli strumenti ultilizzati per l'analisi sono stati addestrati sulla treebank *gold* di IUDT descritta nel precedente capitolo;
- **Passo 2**: creazione di un modello statistico tramite l'applicazione di un algoritmo (LISCA) che sfrutta un preciso set di *feature*s linguistiche estratte probabilisticamente dal corpus automaticamente parsato nella fase 1;
- **Passo 3**: il calcolo di un punteggio di prototipicità (o plausibilità) per ogni relazione di dipendenza presente all'interno della treebank *gold* (utilizzata nella fase 1 di addestramento) utilizzando il modello statistico creato a partire dal corpus giornalistico durante la fase 2.

Il corpus analizzato da LISCA è stato suddiviso in 10 fasce corrispondenti al 10% dell'interno corpus per osservare quali particolari fenomeni linguistici si manifestassero in relazione al decrescere del grado di prototipicità ad essi attribuito.

Il *focus* su cui si sono concentrate le nostre analisi è stato il trattamento da parte dell'algoritmo delle parts-of-speech e delle relazioni di dipendenza, in relazione ad alcuni tra i principali parametri linguistici solitamente utilizzati per la definizione della complessità delle strutture linguistiche: in particolare è stato fatto riferimento alla nozione di marcatezza e alla lunghezza delle relazioni di dipendenza.

In generale è stato possibile notare un forte parallelimso tra le nozioni teoriche di complessità e le analisi realizzate da LISCA. Se da una parte abbiamo visto come le costruzioni sintattiche più prototipiche per l'algoritmo fossero anche quelle meno marcate, in quanto meno ambigue e più

facilmente prevedibili, dall'altra, è stata osservata una forte relazione di interdipendenza tra la complessità computazionale e la marcatezza linguistica: le strutture sintattiche individuate come più complesse e dunque poco prototipiche sono state quelle che a livello morfosintattico e sintattico si distinguono per un maggior grado di marcatezza.

Appendice

1 Relazioni di dipendenza di UD (A-Z)

acl	proposizione implicita od esplicita che modifica un elemento nominale
advcl	proposizione che modifica un verbo o un altro predicato (aggettivo, ecc), come un modificatore
advmod	avverbio o frase avverbiale che serve a modificare il significato della parola
amod	frase aggettivale che serve a modificare il significato del sostantivo
appos	apposizione, elemento nominale immediatamente dopo il primo sostantivo che serve a definire o modificare tale sostantivo. Comprende esempi tra parentesi e abbreviazioni
aux	verbi ausiliari e modali
auxpass	verbi ausiliari e modali al passivo
case	marcatura del caso, relazione utilizzata per qualsiasi elemento trattato come una parola sintattica separata (comprese preposizioni, postposizioni, e clitici). Gli elementi che marcano il caso vengono trattati come dipendenti del sostantivo o delle frasi a cui sono attaccati o che introducono
cc	congiunzione coordinante, relazione tra la prima congiunzione e la congiunzione di coordinazione che delimita un altro elemento coordinato
ccomp	complemento frasale che funge da argomento nucleare
compound	composto
conj	relazione tra due elementi collegati da una congiunzione, come *e*, *o*, ecc. La congiunzione viene trattata asimmetricamente: la testa della relazione è la prima congiunzione e tutti gli altri elementi coordinati dipendono tramite la relazione *conj*
cop	copula
csubj	soggetto sintattico frasale di una frase
csubjpass	soggetto sintattico frasale di una frase passiva
dep	dipendenza non specificata, relazione utilizzata quando un sistema non è in grado di determinare una relazione di dipendenza più precisa tra due parole
det	Determinante, relazione che intercorre tra una testa nominale e il suo determinante

discourse	elemento del discorso, relazione utilizzata per le interiezioni ed altre particelle del discorso
dislocated	elementi dislocati, relazione utilizzata per elementi pre- o post- posti che non soddisfano le solite relazioni grammaticali di base di una frase
dobj	oggetto diretto
expl	elementi nominali espletivi o pleonastici
foreign	parole in lingua straniera
goeswith	*va-con*, relazione che collega due parti di una parola che risultano erroneamente separate
iobj	oggetto indiretto
list	relazione utilizzata per liste di elementi comparabili. Nelle liste con più di due elementi, tutti gli elementi della lista dovrebbero modificare il primo
mark	marcatore, parola che introduce una frase subordinata esplicita
mwe	espressioni multi-parola, relazione utilizzata per certe fisse grammaticalizzate che si comportano come parole piene o avverbi
name	nome, relazione utilizzata per i nomi propri costituiti da più elementi nominali
neg	negazione, relazione tra una elemento di negazione e la parola che modifica
nmod	modificatore nominale, può dipendere sia da un altro nome, sia da un predicato
nsubj	soggetto nominale
nsubjpass	soggetto nominale passivo
nummod	modificatore numerico, qualsiasi numero che modifica il significato del nome con una quantità
parataxis	paratassi, relazione tra una parola (spesso il predicato principale di una frase) e altri elementi, come un'esplicativa o una frase dopo i ":" o un ";", affiancati senza alcuna esplicita coordinazione o subordinazione con la parola testa
punct	punteggiatura. Poiché *punct* non è una relazione di dipendenza normale, i consueti criteri di determinazione della testa non vengono applicati. Invece, vengono utilizzati i seguenti principi: • Un segno di punteggiatura che separa unità coordinate è collegato al primo elemento congiunto • Un segno di punteggiatura che precede o segue un'unità subordinata è collegato a questa unità. • All'interno dell'unità in questione, un segno di punteggiatura è collegato al nodo più alto che conserva la proiettività.

	• Segni di punteggiatura accoppiati (citazioni e staffe) dovrebbero essere collegati alla stessa parola, a meno che creino non proiettività. Questa parola è di solito la testa della frase racchiusa tra la doppia punteggiatura.
remnant	relazione utilizzata per fornire un trattamento soddisfacente dell'ellissi
reparandum	relazione usata per indicare disfluenze in un discorso di riparazione. La disfluenza è il dipendente della riparazione
root	relazione che punta alla radice della frase
vocative	relazione utilizzata per contrassegnare il partecipante a cui ci si rivolge in un dialogo (comune nelle conversazioni, e-mail e gruppi di informazione). La relazione collega il nome del destinatario alla sua frase ospite.
xcomp	proposizione subordinata senza soggetto personale con controllo obbligatorio della proposizione principale.

Bibliografia

Attardi, G. (2006). Experiments with a Multilanguage non-projective dependency parser. In *Proc. of the Tenth CoNLL*.

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).

Berruto, G.; Cerruti, M. (1997). *La linguistica. Un corso introduttivo*. Torino, Utet, pp.7–32.

Bosco, C.; Lombardo, V.; Lesmo, L.; Vassallo, D. (2000). Building a treebank for italian: a data-driven annotation schema. In *Proceedings of LREC 2000*, Athens, Greece.

Bosco, C.; Montemagni, S.; Simi, M. (2012). Harmonization and Merging of two Italian Dependency Treebanks, Workshop on Merging of Language Resources, in *Proceedings of LREC 2012*, Workshop on Language Resource Merging, Instanbul, May 2012, ELRA, pp. 23-30.

Bosco, C.; Montemagni, S.; Simi, M. (2013). Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse* (LAW VII & ID at ACL-2013), Sofia, Bulgaria, August 8-9, pp. 61-69.

Bosco, C.; Dell'Orletta, F.; Montemagni, S.; Sanguintetti, M.; Simi, M. (2014). The Evalita 2014 Dependency Parsing task, *CLiC-it 2014 and EVALITA 2014 Proceedings*, Pisa University Press, ISBN/EAN: 978-886741-472-7, pp. 1-8.

Bosco, C.; Montemagni, S.; Simi, M. (2014). Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. 2014. In *Proceedings of LREC 2014*, ELRA, pp. 83–90.

Bosco C.; Montemagni S.; Simi M. (2015). Harmonizing and merging Italian treebanks: Towards a merged Italian dependency treebank and beyond, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project* 589: 3-23, Springer International Publishing, CH-6330 Cham (ZG) (CHE).

Carroll, J. (2000). Statistical parsing. In Dale, R., Moisl, H. and Somers, H. (eds), *Handbook of Natural Language Processing*, Marcel Dekker, pp. 525–543.

Covington, M. A. (1990). Parsing discontinuous constituents in dependency grammar. *Computational Linguistics* 16, pp 234–236.

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. *Pro- ceedings of the 39th Annual ACM Southeast Conference*, pp. 95–102.

Das, D.; McDonald, R.; Petrov, S.(2012). A Universal Part-of-Speech Tagset, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12).

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.; Montemagni, S. (2012). Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico. In: S. Ferreri (a cura di), Linguistica Educativa. Atti del XLIV Congresso Internazionale di Studi della SLI, Roma, Bulzoni Editore, pp. 343-359.

Dell'Orletta, F.; Montemagni, S.; Venturi, G. (2013). Linguistically-driven selection of correct arcs for dependency parsing, *Computación y Sistemas,* 17(2), pp. 125-136.

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

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. «Cognition», 68(1), pp. 1-76.

Greenberg, J. H. (1962). *Some universals of grammar with particular reference to the order of meaningful elements*, in Id. (edited by), *Universals of language.* Report of a conference held at Dobbs Ferry, N.Y. (April 13-15, 1961), Cambridge (Mass.), The MIT Press, pp. 73-113.

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

Hudson, R. A. (1990). *English Word Grammar*. Blackwell.

Lin, D. (1996). On the structural complexity of natural language sentences. In: Proceedings of COLING 1996, pp. 729–733.

Mel'cˇuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Montemagni, S.; Simi, M. (2007). The Italian dependency annotated corpus developed for the CoNLL–2007 shared task. Technical report, ILC–CNR.

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

Nikula, H. (1986). *Dependensgrammatik*. Liber.

Nivre, J. (2005). Dependency grammar and dependency parsing. Technical Report MSI 05133, Växjö University, School of Mathematics and Systems Engineering.

Nivre, J. (2006). Two strategies for text parsing. In Mickael Suominen, Antti Arppe, Anu Airola, Orvoki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari, K. Pitkänen, and and Kaius Sinnemäki, editors, *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday*, pages 440-448.A special supplement to SKY Journal of Linguistics 19

Nivre, J. (2015). Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3-16. Springer International Publishing Switzerland.

Tesnière, L. (1959). *Éleménts de syntaxe structurale*, Klincksieck, trad. it. (2001), *Elementi di sintassi strutturale*. A cura di Germano Proverbio e Anna Trocini Cerrina. Torino, Rosenberg & Sellier.

Treccani (2011). voce "*Ordine degli elementi"*. http://www.treccani.it/enciclopedia/ ordine-degli-elementi_(Enciclopedia_dell'Italiano)/. Ultima visita: 1/1/2016.

Zwicky, A. M. (1985). Heads. *Journal of Linguistics* 21: 1–29.

Sitografia

http://universaldependencies.org