UNIVERSITÀ DI PISA

DIPARTIMENTO DI FILOLOGIA, LETTERATURA E LINGUISTICA
Corso di Laurea Magistrale in Informatica Umanistica

Studio linguistico-computazionale per l'analisi dei tipi linguistici.

Similarità e differenze nel confronto fra Universal Dependencies Treebanks

TESI DI LAUREA MAGISTRALE

RELATORI
Prof.ssa Simonetta Montemagni
Dott.ssa Giulia Venturi

CONTRORELATORE
Prof.ssa Maria Simi

CANDIDATO
Chiara Alzetta

Anno Accademico 2015-2016

Indice

Capitolo 1

Introduzione

Nel suo discorso di ringraziamento per il *Lifetime Achievement Award* del 2016, conferito dall'*Association for Computational Linguistics* a coloro che maggiormente si sono distinti nel corso della loro carriera nel campo della linguistica computazionale, Joan Bresnan tira le fila del suo percorso di ricerca e riflette su alcuni temi principali che l'hanno accompagnata nel corso della sua carriera. In un'efficace similitudine, che ispira il titolo del suo discorso, dopo aver suddiviso gli approcci all'analisi linguistica in due categorie, paragona una di esse ad un giardino, e l'altra a rovi selvatici.

Nella descrizione che ne fa, il "giardino" si occupa dei dati "coltivati", ovvero quelli dedotti grazie al lavoro dei linguisti, mentre i "rovi" rappresentano le produzioni spontanee dei parlanti, che non sempre sottostanno a tutte le regole della grammatica. Questa suddivisione fa evidentemente eco ad una distinzione già nota in linguistica, formalizzata per la prima volta da Ferdinand De Saussure e rivista dopo circa mezzo secolo da Noam Chomsky: la distinzione fra *langue* e *parole* (o *competence* e *performance*, secondo la nomenclatura Chomskiana). La *langue* riguarda i principi teorici del linguaggio, mentre la *parole* rappresenta le istanze d'uso di tali principi. Chomsky rielabora la *competence* definendola come la conoscenza condivisa da tutti i parlanti di una certa lingua, mentre *performace* è l'uso effettivo del linguaggio in situazioni concrete. L'impatto di una tale bipartizione e delle definizioni dei due linguisti è stato molto significativo, tanto che è forse proprio grazie a ciò che la linguistica tradizionale ha dato il via ad un importante processo di ristrutturazione.

Nel corso del XX secolo infatti è stato dato molto più rilievo al fatto che le produzioni linguistiche spontanee di un parlante spesso sfuggono il controllo o ridefiniscono i

contorni delle grammatiche e delle teorie linguistiche tradizionali. I linguisti hanno quindi iniziato a dedicare parte della loro attenzione anche alla ricerca di sempre nuove tecniche e metodologie che consentissero di ricavare quante più informazioni possibili dalle *performance*. L'obiettivo fondamentale di tali ricerche è spesso quello di inferire le capacità della mente umana per quanto riguarda il linguaggio direttamente a partire dalle strutture attraverso cui esso si realizza. Per questi scopi, alcuni affrontano il problema ricercando i fondamenti biologici che permettono all'essere umano (e solo ad esso) di comunicare attraverso un linguaggio che si articola su più livelli; altri si sono concentrati sullo studio delle regolarità della lingua cercando di estrarle in maniera automatica a partire da grandi quantità di dati linguistici. La conseguenza naturale è stata che la linguistica ha aperto le porte anche ad altre discipline, che con essa si sono intrecciate mettendo a disposizione competenze e metodologie.

L'analisi automatica dei dati linguistici, per esempio, è una metodologia che unisce linguistica e informatica. Questo fortunato matrimonio si è presto rivelato fruttuoso, tanto che può vantare ormai una lunga lista di possibili campi applicativi: la linguistica computazionale e l'NLP (acronimo di *Natural Language Processing*, ovvero Elaborazione del Linguaggio Naturale), così si chiamano le discipline che combinano insieme informatica e linguistica, sfruttano metodi informatici per estrarre dati linguistici e sono oggi campi di ricerca molto attivi, tanto che vantano vivaci comunità non solo all'estero ma anche in Italia. Dalla loro nascita ad oggi in molti casi hanno dimostrato la loro utilità sia nello studio del linguaggio spontaneo che per quanto riguarda le teorie linguistiche classiche; si può ricordare qui per esempio lo sviluppo di sistemi in grado di discriminare le costruzioni grammaticali da quelle non grammaticali, o ancora di quelli che trattano l'uso reale della lingua e le sue deviazioni rispetto alla norma. La linguistica ha potuto inoltre beneficiare dello sviluppo tecnologico nel campo dell'informatica anche per quanto riguarda la grande quantità di dati che al giorno d'oggi possono essere processati: più dati sono sinonimo di più informazioni, che consentono quindi di fare inferenze che si avvicinano maggiormente al reale comportamento della lingua. *More data is better data.*

Inserendosi nel filone della linguistica computazionale e degli studi sulle produzioni spontanee del linguaggio, questa tesi intende porsi come un'ulteriore prova, se fosse necessario, del grande vantaggio che si può ottenere dalla commistione di due discipline apparentemente molto distanti. In particolare, si sfrutteranno algoritmi di analisi linguistica per ricercare le regolarità e irregolarità del linguaggio in tre lingue

diverse, in termini di similarità che le accomunano o differenze che le distinguono. Si è sperimentato l'uso di tali informazioni in uno studio multilingue per ottenere una distinzione delle lingue e dei loro modelli di comportamento su base tipologica e per rintracciare le regolarità universali del linguaggio, valide per tutte o solo per un sottoinsieme delle lingue del confronto. La ricerca delle irregolarità si è concentrata sull'estrazione da corpora annotati delle costruzioni non canoniche, che possono essere tali per più ragioni e su molteplici livelli di analisi. Scopo dell'elaborato sarà dunque dare una definizione delle possibili costruzioni canoniche e non canoniche e mostrarne degli esempi direttamente estratti dai dati linguistici analizzati.

Per rispondere a tali questioni ci si è affidati alla linguistica teorica, in particolare appunto alla linguistica tipologica e alla teoria degli Universali Linguistici. La prima giustifica le similarità e le differenze che si realizzano fra le lingue con l'appartenenza al medesimo gruppo linguistico, che accomuna quelle dotate delle stesse proprietà. La teoria degli universali invece si dedica alla ricerca dei principi universalmente validi per le lingue del mondo nel tentativo di dimostrare che condividono alcune caratteristiche profonde che ne dimostrano la natura biologicamente determinata. Per rintracciare i casi di costruzioni devianti dalla norma, ai principi delle precedenti teorie, che fra di loro presentano importanti punti di contatto, sono stati affiancati gli studi di marcatezza linguistica, che rintracciano quali costruzioni si allontanano maggiormente dalla canonicità di una lingua, ovvero assumono comportamenti inusuali seppur grammaticali.

L'analisi è stata condotta su tre lingue: inglese, italiano e spagnolo. Sebbene si tratti di tre lingue tipologicamente simili, presentano caratteristiche e origini piuttosto diverse. La scelta di tali lingue è stata dettata sia da ragioni di interesse scientifico che da ragioni di ordine pratico. La lingua inglese è quella per cui sono stati messi a punto un numero maggiore di strumenti di analisi linguistica da parte dell'NLP, che ottengono per altro i più alti punteggi di accuratezza. Volendo operare un confronto tipologico con l'italiano, lo spagnolo è stato scelto di conseguenza per poter verificare che le differenze emerse fra le prime due lingue rispecchiassero le proprietà dei corrispondenti tipi linguistici.

Le risorse linguistiche annotate, chiamate anche treebank, che rendono esplicite le informazioni grammaticali contenute nei testi, se rispettano uno standard di rappresentazione sintattica universale, creano i presupposti per realizzare efficacemente studi multilingua come quello presente. Proprio basandosi sui principi della linguistica tipologica e degli Universali Linguistici, negli ultimi anni è stato avviato il progetto

Universal Dependencies (UD). Difatti, pur mettendo a disposizione treebank monolingue, il confronto fra esse è reso possibile da uno standard di annotazione pensato in modo tale da poter rappresentare qualsiasi fenomeno linguistico in maniera coerente, indipendentemente dalle differenze lessicali. Il progetto si pone come obiettivo infatti quello di rappresentare le informazioni morfologiche e sintattiche che, stando alle teorie linguistiche citate, fanno riferimento a proprietà profonde del linguaggio e nella loro forma astratta sono uguali in qualsiasi lingua.

La rappresentazione della morfo-sintassi attraverso il formalismo di UD, come si vedrà anch'esso basato su principi linguistici, consente di utilizzare metodologie e strumenti linguistico-computazionali per l'analisi delle informazioni grammaticali contenute nelle treebank. I parallelismi che si osservano fra le diverse risorse infatti possono essere sfruttati, fra gli altri scopi, per lo studio dei tipi linguistici. Ciò che si cerca di dimostrare attraverso questa ricerca è che il metodo linguistico-computazionale può essere effettivamente utile nel rintracciare le principali similarità (o differenze) fra lingue tipologicamente più o meno vicine; inoltre si vuole mettere in luce come ciò possa portare vantaggi e novità nella teoria linguistica tradizionale, aprendo anche nuovi fronti di ricerca.

Sebbene l'obiettivo iniziale fosse raggiungere tali fini per mezzo di una analisi orientata alla ricerca delle costruzioni tipologicamente marcate attraverso il confronto parallelo fra treebank, di fatto i risultati a cui si è giunti sono molto più ricchi. I comportamenti delle lingue subiscono l'influenza del tipo cui appartengono, per cui le informazioni estratte a partire dalle risorse annotate linguisticamente sono state usate sia per verificare l'efficacia del metodo adottato nell'individuare i tipi, sia per cercare nuove similarità non ancora attestate fra le lingue. Inoltre, non secondarie, si sono cercate e rintracciate anche delle differenze fra treebank, alcune dettate dalle proprietà e caratteristiche della lingua, altre dallo schema di annotazione che esse adoperano. Ai risultati linguistici infatti si sono poi di fatto affiancati anche dati relativi al contenuto delle treebank e informazioni riguardanti criteri adoperati da ciascuna lingua nello svolgere l'annotazione. Pur rispettando tutte uno schema di annotazione universale come quello di UD, si rintracciano delle differenze legate a scelte individuali che ostacolano la ricerca.

Per giungere a questi risultati è stato utilizzato LISCA (acronimo di *LInguistically driven Selection of Correct Arcs*). Si tratta di uno algoritmo che assegna ad ogni relazione grammaticale un punteggio di plausibilità sulla base del contesto strutturale

e linguistico della frase in cui l'arco che la rappresenta è inserito. L'algoritmo, sviluppato presso l'Istituto di Linguistica Computazionale del CNR di Pisa, può essere eseguito sia su corpora annotati automaticamente da strumenti detti *parser,* che su risorse linguistiche annotate manualmente, come è stato fatto in questo caso. La scelta di utilizzare uno strumento come questo su treebank *gold*, ovvero idealmente prive di errori, piuttosto che su corpora annotati tramite parser è stata guidata dal desiderio di rintracciare non tanto informazioni riguardanti lo schema di annotazione, quanto piuttosto i dati linguistici che dall'annotazione possono emergere. I fenomeni della treebank vengono quindi ordinati sulla base del punteggio ottenuto, andando a costituire una "classifica" in cui le relazioni si collocano più in alto o più in basso a seconda della possibilità che hanno di realizzarsi nelle produzioni linguistiche. Nella parte più alta della classifica prenderanno posto le relazioni più semplici e comuni. In fondo sarà possibile trovare le costruzioni più rare e meno probabili, fra le quali si nasconderanno anche gli errori di annotazione. Difatti, sebbene la ricerca sia stata svolta su risorse linguistiche annotate manualmente, di fatto pare che un margine di errore debba sempre essere tenuto in conto. L'algoritmo di LISCA si è rivelato utile anche per individuare quelle relazioni che necessitano di revisione.

L'elaborato presenterà innanzitutto la storia e principi fondamentali della linguistica tipologica e della teoria degli Universali Linguistici, ad essa fortemente connessa, passando poi ad illustrare la marcatezza come fenomeno linguistico e cosa si intende per marcatezza tipologica (capitolo 2). Nel capitolo 3 si tratterà del formalismo della grammatica a dipendenze e del progetto Universal Dependencies, che sfrutta tale formalismo nel mettere a disposizione delle risorse linguistiche annotate manualmente; verrà inoltre fatto un breve accenno alle tecniche di parsing automatico e di annotazione manuale morfo-sintattica. Il capitolo 4 è dedicato alla descrizione degli strumenti e dei dati utilizzati nello studio, mentre il capitolo 5 descrive nel dettaglio i risultati ottenuti. Al capitolo 6 sono infine dedicate le conclusioni che contengono le nuove prospettive di ricerca che si aprono grazie a questo studio.

Capitolo 2

Linguistica tipologica e fenomeni linguistici

Ogni lingua del mondo è diversa dalle altre. Anche le lingue che vengono definite simili di fatto hanno delle differenze sostanziali che non permettono a coloro che ne parlano una di sapersi esprimere correttamente anche nell'altra senza prima passare attraverso una fase di apprendimento delle regole grammaticali. Tuttavia, anche senza conoscerne nel dettaglio la struttura, è possibile rintracciare delle similarità fra alcune lingue, mentre altre paiono talmente diverse che sembrano non avere nulla in comune. Tali similarità vengono intuitivamente ricondotte alla vicinanza geografica e al contatto che due o più lingue hanno avuto nel corso della storia, influenzandosi a vicenda, oppure ancora alla discendenza da una lingua antenata comune di cui hanno tutte mantenuto certe proprietà. Tutte le lingue infatti possono essere ricondotte a determinate famiglie linguistiche, fra i cui membri si possono rintracciare delle analogie. Tuttavia possono esserci anche altre ragioni per cui due o più lingue presentano delle affinità di comportamento [Moravcsik, 2013]. La disciplina che si occupa della ricerca e dello studio di tali similarità è la linguistica tipologica.

2.1 Linguistica tipologica

Il termine "tipologia" nasce in biologia come sinonimo di "tassonomia", "classificazione" [Croft, 2002]. Il primo ad usarlo in linguistica fu von der Gabelentz [1901], il quale sostenne che una classificazione delle lingue su base genealogica non può essere

equiparabile ad una basata sulle loro caratteristiche grammaticali. Come in biologia, la tipologia si ricava attraverso la ricerca di modelli ricorrenti; la linguistica tipologica si occupa dunque dello studio di differenze e similarità strutturali fra sistemi linguistici [Velupillai, 2012]. Fattori storici, "genetici" e ambientali possono influire sulla struttura di una lingua, ma la classificazione per tipi è quella che permette di avere dei raggruppamenti sulla base della struttura.

In letteratura si possono rintracciare numerose definizioni di linguistica tipologica. Dressler [1987: 470] per esempio sostiene che "per linguistica tipologica si intende la definizione di categorie linguistiche generali che fungano da base per la classificazione delle lingue in tipi indipendentemente dalle loro origini storiche. In particolare fare linguistica tipologica significa generalizzare le tipologie delle lingue sulla base di similarità e dissimilarità nella loro struttura linguistica"[1]. In questa definizione nessuna componente, ad eccezione della vera e propria struttura, serve nella delineazione di un tipo. Già nel 1921 Sapir era giunto ad una simile conclusione: nel libro *Linguaggio: Introduzione alla linguistica* afferma che le classificazioni delle lingue si possono fare solo sulla base della "natura dei concetti espressi dal linguaggio" poiché ciascuna lingua può esprimere gli stessi concetti mediante strategie morfologiche e sintattiche diverse [Sapir, 1921: 136]. Già era chiaro dunque che gli elementi che accomunano le lingue risiedono ad un livello profondo, mentre le differenze sono superficiali.

La vera svolta che ha conferito alla disciplina i connotati moderni è però arrivata solo con Greenberg e la sua analisi inter-linguistica sull'ordine delle parole nella frase [1963]. Greenberg si ispirava ai linguisti della Scuola di Praga, e in particolare a Jakobson, che già si era dedicato allo studio del modo in cui si combinano fra loro diverse proprietà della lingua [Jakobson, 1932]. Allo stesso tempo a Greenberg va anche il merito di aver cercato di stabilire dei metodi quantitativi per lo studio delle tipologie linguistiche che permettessero alla disciplina di rispettare gli standard della ricerca scientifica [Greenberg, 1960].

Nel tentativo di sintetizzare le diverse posizioni dei linguisti sugli approcci e gli scopi della linguistica tipologica, Croft [2002] distingue due generiche macro-definizioni in cui si possono far ricadere tutte le altre. La meno specifica delle due fa riferimento alla classificazione dei tipi strutturali che si manifestano nelle lingue. In questa

[1]Testo originale: "*By linguistic typology we mean the determination of general linguistic categories as a basis to classify languages into types regardless of their historical origin [....] In particular, linguistic typology means generalizing typology of languages according to the similarity or dissimilarity in their linguistic structure*".

prospettiva, ogni lingua ricade all'interno di un tipo e il compito della disciplina è rintracciare e definire tutti i tipi esistenti. La seconda definizione è solo apparentemente molto simile alla precedente: il compito della linguistica tipologica è riconoscere e studiare tutti i modelli di comportamento che si ripropongono sistematicamente fra le lingue. Si fa riferimento in questa definizione agli universali linguistici, ovvero quegli atteggiamenti che le lingue assumono che, anche se non sono comuni ad ogni lingua osservabile, comunque non vengono contraddetti da nessuna [Greenberg, 1960]. In ogni caso entrambe le definizioni sottolineano il fatto che gli studi tipologici debbano essere condotti operando un confronto fra lingue [Ramat, 1987].

Gli studi tipologici vengono spesso criticati perché risultano meramente descrittivi e tassonomici. Questa accusa può essere valida per il primo dei due approcci presentati da Croft [2002], ma la ricerca degli universali è tutt'altro che un processo descrittivo: si tratta di una indagine complessa che prende in considerazione numerose componenti che emergono dal confronto fra lingue e che possono addirittura fornire numerose informazioni sulla natura stessa del linguaggio [Comrie, 1993].

2.2 I tipi linguistici

Qualsiasi livello del linguaggio può essere il punto di partenza per un confronto tipologico: fonologia, morfologia, sintassi, lessico, pragmatica, semantica , etc. [Velupillai, 2012]. L'obiettivo fondamentale di una qualsiasi indagine tipologica è infatti trovare risposta alle seguenti tre domande [Bickel, 2007]:

1. Che cosa? Ovvero quali fenomeni linguistici si possono osservare nelle lingue naturali;
2. Dove? Quali lingue manifestano tali fenomeni;
3. Perché? Per quale ragione certi modelli si propongono sostanzialmente uguali in più lingue.

I livelli linguistici che solitamente vengono presi in considerazione e che vanno a delineare i tipi più importanti sono la morfologia e la sintassi. In questa sezione si guarderà prima ai diversi tipi morfologici possibili e successivamente verranno descritti i tipi sintattici.

2.2.1 Tipi morfologici

La morfologia descrive quali elementi sono necessari per costruire parole ben formate in una data lingua. Questi elementi sono per l'appunto i morfemi, le unità minime di una lingua dotate di significato, che si combinano fra loro rispettando regole ben precise [Moravcsik, 2013]. Osservando queste regole si può notare che certe lingue sono dotate di poca o nessuna morfologia flessionale, mente altre ne sono ricche. Questa prima informazione permette di individuare due tipi linguistici opposti: le lingue analitiche e le lingue sintetiche. Sono analitiche le lingue che codificano informazioni sintattiche attraverso particelle, verbi ausiliari, etc.; viceversa sono sintetiche quelle che usano esprimere tali rapporti per mezzo di desinenze e variazioni tematiche. Nelle prime il rapporto fra la parola e i morfemi che la compongono, detto indice di sintesi, è 1:1, mentre nelle seconde l'indice di sintesi 1:molti [Schlegel, 1818]. Un esempio evidente si può osservare nella differenza fra latino e italiano, dove *patris* in latino corrisponde a *del padre* in italiano: il morfema *-is* veicola un'informazione sintattica che in italiano invece è rappresentata dalla preposizione articolata *del*. A queste si aggiungono le lingue polisintetiche che non solo fanno uso di un gran numero di morfemi, ma possono anche avere più di una radice lessicale all'interno della stessa parola.

La formulazione classica di tipologie morfologiche fatta da Schleicher a metà del XIX secolo individuava anche altri tre tipi, dislocandoli lungo una scala lineare immaginaria che ha ad un estremo le lingue isolanti, all'altro le flessive e in mezzo fra le due le agglutinanti [Schleicher, 1983].

Secondo questa classificazione, le lingue isolanti sono quelle che non si servono di affissi morfologici e hanno poca o nulla morfologia derivazionale. Le lingue di questo gruppo sono anche lingue analitiche poiché il loro indice di sintesi è generalmente 1:1. Inoltre, oltre che monomorfemiche, queste lingue tendono ad essere anche monosillabiche.

Agglutinare invece etimologicamente significa "incollare assieme", e infatti le lingue agglutinanti si servono di un morfema diverso per ogni informazione grammaticale e li concatenano insieme per formare le parole, che come risultato hanno una struttura interna piuttosto complessa. Il loro indice di sintesi è infatti pari o superiore a 1:3. Dal momento che ogni morfema tende ad avere una valore univoco e a marcare un'unica categoria grammaticale, sono molto rari i fenomeni di omonimia o allomorfia

fra morfemi. In una lingua agglutinante le parole possono essere anche molto lunghe, costituite da una radice lessicale a cui sono attaccati più affissi; difatti quello che in una lingua agglutinante è una sola parola spesso corrisponde ad un sintagma in una lingua flessiva.

Le lingue flessive (o fusive) infatti sintetizzano all'interno di un unico morfema più categorie grammaticali. In questo tipo le parole sono internamente piuttosto complesse, costituite da una base lessicale semplice o derivata (radice) e da uno o più affissi flessionali che spesso sono morfemi cumulativi, veicolando ciascuno più valori grammaticali assieme e assommando diverse funzioni. L'indice di sintesi in questo caso è comunque più semplice rispetto alle lingue agglutinanti, si aggira infatti intorno a 1:2 o 1:3, denotando parole con una struttura meno complessa e composte da un numero minore di morfemi. Viceversa però non sono rari nelle lingue flessive i casi di allomorfia e di fusione, che rendono i morfemi non ben separabili e difficilmente identificabili.

Esempi prototipici di lingue isolanti sono il vietnamita e il cinese mandarino; il turco è una lingua agglutinante; il latino e le lingue romanze sono lingue flessive.

A questi tre tipi fondamentali se ne aggiungono solitamente altri che ne costituiscono dei sottotipi.

Il sottotipo delle lingue flessive è rappresentato dalle lingue introflessive, come l'arabo, in cui i fenomeni di flessione avvengono anche all'interno della radice lessicale. Si parla in questo caso di transfissi vocalici che si inseriscono fra le consonanti della base discontinua della radice. Le lingue polisintetiche invece, a volte chiamate incorporanti, hanno una struttura interna della parola ancora più complessa: l'indice di sintesi è solitamente 1:4 ma, a differenza delle agglutinanti a cui sotto molti aspetti somigliano, in una stessa parola possono apparire anche due o più radici lessicali. Inoltre, come il tipo flessivo, le lingue polisintetiche presentano fenomeni di fusione che rendono poco trasparente la struttura della parola. Un esempio di lingua polisintetica è il groenlandese [Berruto and Cerruti, 2011].

2.2.2 Tipi sintattici

Dal punto di vista sintattico le lingue possono essere classificate sulla base delle categorie grammaticali che presentano: per esempio non tutte le lingue possiedono

gli articoli o li usano allo stesso modo; questa distinzione già permette di creare raggruppamenti. Il criterio fondamentale per la classificazione delle lingue in tipi basati sulla sintassi è l'ordine non marcato nelle frasi dichiarative canoniche dei costituenti principali, ovvero soggetto (S), predicato verbale (V) e oggetto (O). Questi tre elementi possono combinarsi secondo sei ordini possibili:

- Soggetto Verbo Oggetto (SVO);
- Soggetto Oggetto Verbo (SOV);
- Verbo Soggetto Oggetto (VSO);
- Verbo Oggetto Soggetto (VOS);
- Oggetto Soggetto Verbo (OSV);
- Oggetto Verbo Soggetto (OVS).

Sebbene ciascuno di questi ordini possa essere rintracciato in almeno una lingua del mondo, i primi due (SVO e SOV) sono i tipi più comuni, mentre l'ordine OSV appare in pochissimi casi[2]. Le ragioni di questa distribuzione riguardano il fatto che solitamente soggetto e tema della frase coincidono, e nell'ordine lineare dei costituenti informativi ciò di cui si parla viene prima di ciò che si dice al proposito. Inoltre sembrano agire sull'ordine dei costituenti altri due principi, entrambi rispettati dalle lingue SVO e SOV: il principio di precedenza e il principio di adiacenza. Il principio di precedenza prescrive che fra i costituenti nominali quello con maggiore prominenza e priorità logica (il soggetto) debba trovarsi in prima posizione. Il secondo principio invece stabilisce che ci debba essere un rapporto di vicinanza fra verbo e oggetto in virtù della loro stretta relazione sintattico-semantica[3]. Gli altri ordini sarebbero sfavoriti perché non obbediscono a tali principi [Berruto and Cerruti, 2011].

Il fatto che una lingua individui un ordine canonico per i costituenti di una frase non implica che non sia comunque ammesso un certo grado di libertà: in alcune lingue che lo consentono si possono realizzare anche delle variazioni nell'ordine degli

[2]Dryer [1995]ha rintracciato la seguente distribuzione: SOV 497 lingue, SVO 435 lingue, VSO 85 lingue, VOS 26 lingue, OVS 9 lingue, OSV 4 lingue. 172 lingue non presentano un ordine lineare dominante.

[3]La letteratura linguistico-computazionale in materia di rappresentazione sintattica fanno riferimento a questo fenomeno col nome di DDM (*Dependency Distance Minimisation*) [Temperley, 2007].

elementi che veicolano specifici intenti comunicativi e che costituiscono costruzioni con ordine marcato [Payne, 1990].

Dalla posizione dei costituenti fondamentali si è anche cercato di derivare quale potesse essere l'ordine di base degli altri elementi linguistici in costrutti più articolati. Tali regole sono definite attraverso degli universali implicazionali, ovvero dei principi generalmente validi e mai contraddetti che mettono in relazione fra loro le posizioni dei diversi elementi nella frase e nei sintagmi. Sulla base della posizione di oggetto e verbo, per esempio, è possibile definire due tipi linguistici fondamentali: le lingue VO e le lingue OV. Attraverso gli universali implicazionali è stato messo in luce che lingue VO sono lingue postdeterminanti, ovvero rispettano un ordine testa-modificatore che si ripropone in diversi casi: l'aggettivo si colloca dopo il nome, lo stesso il possessivo, l'avverbio dopo il verbo, etc. Il contrario accade per le lingue OV, che invece sono prederminanti.

Si tenga presente tuttavia che tutte le lingue presentano un certo grado di incoerenza tipologica interna: in nessun caso una lingua manifesta le caratteristiche di un unico tipo. Per questa ragione, quando si cerca di classificare le lingue sulla base del tipo linguistico di appartenenza, si sottolinea sempre che ciascuna di esse presenta delle tendenze che coincidono prevalentemente con le caratteristiche di un determinato tipo, ma che a queste se ne accompagnano altre possibilmente tipiche di un'altra categoria.

2.3 Universali linguistici

Lo studio dei tipi linguistici e il tentativo di ricondurre tutte le lingue naturali a categorie sulla base delle caratteristiche comuni che manifestano non può non intrecciarsi con la teoria degli universali linguistici.

Per universale linguistico si intende una proprietà che risulta comune a tutte le lingue del mondo, o ad un gran numero di esse, e che non viene contraddetta da nessuna lingua. L'idea che le lingue naturali possiedano tratti comuni è piuttosto antica e affonda le radici nella filosofia, ma è anche in questo caso con Greenberg negli anni Sessanta che si è potuto assistere ad un vero rinnovamento della teoria. Greenberg si mise a capo di una ricerca il cui scopo era il confronto parallelo di trenta lingue diverse, anche fra loro geneticamente distanti. Il risultato di questo studio fu un

catalogo di tutte le proprietà che le accomunano e la quantità di similarità fra esse risultò talmente alta che non era possibile si trattasse di dati casuali. Greenberg ha dimostrato che le lingue naturali condividono delle caratteristiche profonde, indipendentemente dal loro aspetto superficiale. Oltre a questo a lui va riconosciuto il grande merito di aver formalizzato un approccio allo studio degli universali linguistici: solo l'osservazione di dati provenienti dal più alto numero di lingue possibili e il loro confronto diretto può far emergere le caratteristiche che le accomunano.

Contemporaneamente a Greenberg, Noam Chomsky stava delineando i dettagli della sua Grammatica Generativa, che permette l'indagine degli universali attraverso un approccio più astratto mirato alla ricerca delle strutture formali partendo dall'osservazione di un numero minore di confronti. Gli universali di Chomsky servono prevalentemente a spiegare la natura innata e biologicamente determinata del linguaggio; questa è la ragione per cui lo studio degli universali, più di altre teorie, ha fatto sì che la linguistica si inserisse di fatto fra quelle che sono chiamate scienze cognitive. Gli approcci proposti da Greenberg e Chomsky sono i due metodi più popolari per lo studio delle proprietà comuni del linguaggio. Tuttavia nel corso degli anni le distanze fra le due posizioni si sono assottigliate e ciascun modello, qualunque esso sia, riconosce la necessità di tenere in considerazione i fenomeni grammaticali in una prospettiva multilingua [Mairal and Gil, 2006].

Gli universali linguistici si suddividono in sottogruppi sulla base del comportamento che descrivono.

Una prima distinzione, suggerita da Chomsky, riguarda gli universali sostanziali e formali. Gli universali sostanziali fanno riferimento a quegli elementi linguistici che tutte le lingue necessariamente devono possedere. Per esempio tutte le lingue naturali possiedono le vocali o delle particelle specializzate per rappresentare il ruolo del parlante (io) opposto al ricevente (tu). Gli universali formali invece si riferiscono ai meccanismi che le lingue devono mettere in opera per poter realizzare costruzioni specifiche: per esempio per ottenere una frase interrogativa bisogna compiere certe manipolazioni sulla dichiarativa corrispondente che possono tuttavia essere diverse da lingua a lingua.

Oggetto di maggiore interesse da parte dei linguisti sono stati gli universali implicazionali. Si tratta di un'altra classificazione, trasversale alle due precedenti, che comprende quegli universali che mettono in correlazione più proprietà linguistiche. Non tutte le proprietà sembrano essere influenzate dalle altre, mentre alcune sono

evidentemente più basilari e hanno un forte impatto sugli elementi della frase in cui appaiono [Croft, 2002]. Gli universali implicazionali si costruiscono rispettando uno dei seguenti schemi formali (dove A e B sono due proprietà qualsiasi):

- Le lingue che possiedono A allora possiedono B;
- Le lingue che possiedono A allora non possiedono B;
- Le lingue che non possiedono A allora possiedono B;
- Le lingue che non possiedono A allora non possiedono B.

Se un vasto numero di lingue rispetta una certa norma, formalizzata come qui sopra, e nessuna la contraddice, allora quello che si sta osservando è un universale implicazionale.

Tuttavia, per evitare di costruire degli universali privi di significato, ciascuna regola deve rispettare contemporaneamente almeno tre schemi su quattro. Si veda per esempio uno dei più noti universali implicazionali: se una lingua possiede vocali nasali allora possiede anche vocali non nasali. Questo è un universale noto e afferma che le vocali nasali sono in un qualche modo un'evoluzione delle vocali non nasali, da cui devono derivare. Si consideri invece la seguente affermazione: se una lingua possiede vocali nasali, allora ha anche vocali orali. Di questo universale, che potrebbe essere confuso per vero, esistono di fatto solo due combinazioni effettivamente riscontrabili fra le formalizzazioni proposte: esistono lingue che possiedono sia vocali nasali che orali (A e B), esistono anche lingue che non possiedono vocali nasali ma possiedono vocali orali (non-A e B), ma non esistono lingue che non possiedono nessuna vocale (non-A e non-B). In questi casi quello che si può costruire è un universale non-implicazionale, ovvero che tutte le lingue hanno vocali orali, il quale rende l'universale descritto prima superfluo [Comrie, 1993].

Un'ultima distinzione che solitamente viene fatta distingue gli universali fra assoluti e tendenze. Gli universali assoluti sono quelli che si realizzano senza eccezioni, mentre gli altri descrivono piuttosto delle tendenze non vincolanti.

Uno dei grandi meriti della teoria degli universali linguistici è di aver fatto sì che la linguistica iniziasse ad osservare i fenomeni grammaticali in una prospettiva interlinguistica. È il caso questo dei fenomeni di marcatezza.

2.4 Marcatezza

La marcatezza linguistica è un fenomeno che può realizzarsi a vari livelli di analisi e si riferisce a quelle costruzioni, o quelle proprietà, che si distanziano per una qualche ragione dal comportamento canonico della lingua. Questa definizione, sebbene generica e astratta, cerca di far emergere i tratti fondamentali comuni a tutte le descrizioni di marcatezza che si trovano in letteratura: nel corso degli anni la marcatezza è stata oggetto di varie interpretazioni che hanno posto l'accento su certe caratteristiche o su altre.

La nozione di marcatezza viene formalmente introdotta in linguistica a partire dagli anni trenta. Il termine nasce negli studi di fonetica di Trubetzkoy [1939], ma oggi si ritiene che la sua sia una definizione troppo astratta e generica, tanto più che fa riferimento a tratti distintivi contestuali dei fonemi e non può essere applicata alle proprietà universali della lingua (Haspelmath, 2006 e Andrews, 1990). Ben più fortunata è stata l'interpretazione fornita da Jakobson, il quale non solo analizza i casi di marcatezza come opposizioni binarie asimmetriche fra fenomeni, ma estende la nozione a tutti i possibili livelli di analisi linguistica. Con la pubblicazione nel 1941 del volume *Linguaggio infantile, afasia e leggi fonetiche universali*[4], Jakobson inoltre associa la marcatezza alla componente cognitiva che regola il linguaggio. Partendo dai dati della ricerca medica, l'autore osserva che due fenomeni opposti, come apprendimento della lingua materna nel bambino e perdita progressiva del linguaggio in soggetti afasici, seguono in realtà andamenti specularmente identici: le costruzioni che vengono apprese per ultime sono le prime che il soggetto affetto da afasia tende a perdere, e viceversa. Tale processo non pare essere influenzato dalla lingua dei soggetti e si ripropone in ogni caso costante, sottostando ad un principio di economia linguistica biologicamente determinato.

Dai lavori di Jakobson in poi non si è più potuto parlare di marcatezza osservandola solo sul piano formale, ma tenendo sempre in considerazione anche i casi d'uso della lingua: ad oggi "non marcato" viene spesso utilizzato come sinonimo di semplice, comune, facile da produrre, acquisito più rapidamente, etc. [Hume, 2004]. A questa concezione ha contribuito anche Greenberg, il quale ha dato una propria interpretazione di marcatezza che è stata poi ribattezzata da Zwicky come "marcatezza statistica". Greenberg sostiene che la proprietà veramente fondamentale di un fenomeno marcato è la bassa frequenza d'uso che la contraddistingue. Questa proprietà

[4]Titolo originale: *Kindersprache, Aphasie und allgemeine Lautgesetze.*

deriva direttamente dalla definizione di marcatezza come relazione asimmetrica, già utilizzata da Jakobson: l'asimmetria si realizza come diversa frequenza d'uso nelle produzioni dei parlanti madrelingua. Lo studio dei fenomeni della lingua attraverso l'osservazione della loro frequenza d'utilizzo porta con sé un indubbio vantaggio metodologico: l'estrazione di dati statistici dai testi può essere realizzata su qualsiasi lingua del mondo senza restrizioni, rendendo i risultati ottenuti sulle diverse lingue confrontabili fra loro [Kucera, 1982]. In questa prospettiva si sono sviluppati i successivi studi di marcatezza tipologica.

La definizione di Greenberg fa eco al concetto di predicibilità: "la marcatezza è da considerarsi come una nozione probabilistica in cui la predicibilità è positivamente correlata con la non marcatezza"[5] [Hume, 2004: 194]. La predicibilità è un concetto che fortemente si lega a quello di esperienza e frequenza di utilizzo e che quindi può essere calcolato attraverso l'osservazione delle produzioni spontanee dei parlanti: le forme più predicibili sono quelle non marcate, viceversa quelle marcate sono le meno predicibili.

La marcatezza viene comunque tradizionalmente intesa come un fenomeno interno alla lingua: una costruzione è marcata, comunque si intenda il termine, in rapporto ad altre costruzioni equivalenti nello stesso idioma. Intendere invece la costruzione marcata come la meno probabile in termini quantitativi ci permette di osservare il fenomeno in una prospettiva multilingua.

"Una struttura X è tipologicamente marcata rispetto ad una struttura Y se tutte le lingue che presentano X hanno anche Y." [Gundel et al., 1986].

Questa definizione identifica la marcatezza tipologica come una proprietà asimmetrica, transitiva ma non riflessiva, propria delle lingue naturali, tale che la presenza di una certa struttura ne implica un'altra, ma non viceversa [Eckman, 2008: 97]. Sono stati infatti svolti esperimenti sui processi di apprendimento di lingue seconde (L2, in opposizione alla lingua materna a cui si fa riferimento come L1). Sebbene le costruzioni marcate siano acquisite con difficoltà dagli apprendenti L2, indipendentemente dalla lingua materna di partenza, il tipo linguistico di L1 e L2 ha un forte impatto sul grado di difficoltà che il soggetto riscontra. L'apprendimento di forme marcate in una L2 tipologicamente molto lontana dalla L1 di partenza risulta più

[5]Originale: "*Markedness is best considered a probabilistic notion with predictability positively correlated with unmarkedness*".

complesso per il soggetto rispetto a quanto non accada per L2 tipologicamente simili alla propria L1 [Eckman, 2008: 5].

Si sarà notato che, oltre alla frequenza, si è fatto riferimento alle costruzioni marcate come costruzioni complesse. La complessità di una frase può derivare da diversi fattori e può essere osservata su più piani: complessità lessicale, sintattica, formale, etc. [Gibson, 1998]. "Da un punto di vista formale, le forme marcate, in confronto alle corrispondenti forme non marcate, sono morfologicamente più complesse e meno lessicalizzate, [...] meno neutrali in termini di registro. Dal punto di vista del significato, tali forme suggeriscono un significato aggiuntivo o una connotazione assente nelle corrispondenti forme non marcate." [Levinson, 2000][6]. Mentre una frase può avere diversi gradi di complessità, una costruzione marcata è complessa di natura, sia che la si intenda complessa perché semanticamente più connotata, che perché morfologicamente più ricca, o ancora perché inattesa. Il fenomeno si può osservare quindi anche in un'altra prospettiva: tanto più un certo fenomeno linguistico è diverso dallo standard, tanto più è complesso e tanto meno verrà usato.

2.5 Grammatica Universale

Come si sarà notato, la descrizione e i principi della linguistica tipologica e della teoria degli universali linguistici spesso coincidono. Si tratta infatti di due discipline sorelle fra cui è difficile delineare confini netti: nel caso della tipologia linguistica vengono studiate le restrizioni nelle variazioni inter-linguistiche, gli universali cercano invece di darne una definizione. Per questa ragione la ricerca tipologica può essere vista come il metodo principale per far emergere gli universali linguistici [Comrie, 1993].

Parallelamente a queste due teorie è però possibile osservarne una terza, che con esse condivide i principi sostanziali: la Grammatica Universale (UG, *Universal Grammar*). Il nome stesso è esplicativo del concetto fondamentale che sta alla base di questa teoria; ovvero che nella sua forma più sostanziale la grammatica è unica per tutte le lingue, anche se accidentalmente varia [Bacon, 1902]. Questa idea, già piuttosto antica, è stata ripresa da Chomsky, il vero padre della UG, e dopo di lui

[6]Testo originale: *"On the formal side, marked forms, in comparison to corresponding unmarked forms, are more morphologically complex and less lexicalized, [...] less neutral in register. On the meaning side, such forms suggest some additional meaning or connotation absent from the corresponding unmarked forms."*

ha avuto grande fortuna in vari ambiti quali psicologia, studio dell'acquisizione ed elaborazione computazionale del linguaggio.

Il focus dei lavori di Chomsky è sempre stata la mente umana e il modo in cui elabora l'informazione linguistica. Difatti la UG nasce proprio nel tentativo di spiegare secondo quali principi gli umani apprendono il linguaggio e che cosa sa un individuo quando conosce una lingua. Per Chomsky la Grammatica Universale è una teoria della conoscenza che si occupa delle strutture (necessariamente) interne alla mente umana che collegano i suoni della lingua ai concetti che essa esprime [Chomsky, 1965].

Secondo la famosa teoria dei Principi e Parametri che lo stesso autore ha delineato, apprendere una lingua significa conoscere il modo in cui i principi del linguaggio si applicano ad essa e quali i valori da assegnare ai parametri di modo che le regole della grammatica siano rispettate. Lo scopo fondamentale della teoria della UG è quindi analizzare i fenomeni linguistici tenendo conto contemporaneamente del modo in cui si realizzano, della mente umana e della grammatica [Chomsky and Lasnik, 1993]. Il fatto che alcuni elementi del linguaggio siano radicati nella mente, o ancora meglio siano influenzati dalle strutture mentali, spiega per Chomsky come un bambino riesca ad apprendere rapidamente la lingua materna anche in circostanze di *povertà dello stimolo* [Cook, 1985].

Sebbene la teoria della Grammatica Universale di Chomsky, dopo un iniziale entusiasmo, sia stata anche fortemente criticata dai linguisti e psicolinguisti, soprattutto per quanto riguarda il modo in cui vengono definiti i processi che porterebbero all'apprendimento del linguaggio[7], è tutt'ora fortemente "viva" l'idea che in tutte le lingue siano rintracciabili dei comportamenti *universali*.

2.6 Verso l'approccio linguistico-computazionale

Il linguaggio, nonostante sia il prodotto della creatività umana e quindi per natura variabile e ambiguo, possiede delle regolarità che consentono di indagarlo attraverso metodi quantitativi: la linguistica computazionale e l'NLP si occupano di definire strategie computazionali per l'elaborazione e l'estrazione di informazione contenuta nel linguaggio naturale.

[7]Si vedano, per un approfondimento, gli studi di Tommasello.

Tali informazioni non possono emergere dai dati grazie esclusivamente all'applicazione di strumenti che realizzano indagini statistiche; i risultati devono sempre essere accompagnati dalla loro interpretazione, che solitamente si basa sui principi definiti dalle teorie linguistiche, supportandole, confutandole o delineandone di nuove. Queste discipline si affiancano infatti alla linguistica teorica fornendole dei nuovi strumenti di indagine.

Universali e tipi linguistici, insieme alla UG, sono fra le teorie che più possono beneficiare della combinazione fra linguistica e informatica a cui si è assistito nel corso del XX secolo. Infatti esse interpretano i fenomeni del linguaggio servendosi delle nozioni di distribuzione e probabilità, che sono proprie della statistica. Il già citato fenomeno della marcatezza, per esempio, è stato interpretato più volte come "marcatezza statistica", ovvero in termini distribuzionali. In questa accezione, più un fenomeno è marcato e tanto più è improbabile che si realizzi nel linguaggio, pur essendo di fatto grammaticale. Di conseguenza un fenomeno marcato è quello di cui si osservano poche occorrenze o che deriva da una combinazione di fattori che solitamente non occorrono insieme. La ricerca dei fenomeni di marcatezza all'interno di una risorsa linguistica può quindi essere svolta anche adottando metodologie linguistico-computazionali per intercettare comportamenti statisticamente rilevanti inducendoli a partire da dati grezzi, ovvero testi rappresentativi del registro standard di una lingua. Una volta stabilito, per mezzo di calcoli statistici, quali costruzioni di una lingua si possono considerare standard, i fenomeni marcati saranno quelli che deviano maggiormente da essi.

Grazie alla teoria degli universali e dei tipi linguistici è possibile affrontare la marcatezza anche in una prospettiva multilingua. La teoria degli universali linguistici sostiene che il linguaggio è il prodotto della mente umana e che le lingue sono solamente manifestazioni del linguaggio che differiscono per aspetti superficiali mentre sono accomunate da caratteristiche profonde. La linguistica tipologica afferma inoltre che le lingue possono essere raggruppate sulla base delle differenze superficiali che manifestano. Date queste premesse, la marcatezza risulta come un fenomeno linguistico che si realizza nelle stesse costruzioni per tutte le lingue di un dato tipo e che può essere esplorato attraverso un confronto multilingua. Ed è proprio l'approccio linguistico-computazionale che permette di carpire le informazioni linguistiche contenute nei testi, oppure quali sono i punti di contatto e di distanza maggiore fra due o più lingue.

Bisogna tener presente che, basandosi su modelli statistici, quello che gli strumenti computazionali catturano meglio sono le tendenze delle lingue, ovvero modelli di comportamento che si realizzano, non senza eccezioni, in determinati contesti. Andare ad analizzare ed interpretare i risultati di tali ricerche impone quindi anche di trovare una spiegazione al perché, date certe condizioni, la lingua tendenzialmente si comporta secondo certe regole, quali i casi in cui questo non avviene (fermo restando che la grammaticalità della lingua deve essere comunque rispettata) e per quali ragioni.

Gli strumenti computazionali che consentono di estrarre questo tipo di informazione dai testi scritti in linguaggio naturale hanno di fatto una molteplicità di applicazioni possibili. In questo contesto si è scelto di impiegarli per andare ad indagare la distanza fra lingue, ricavando in questo modo informazioni sulla tipologia di appartenenza e sulle costruzioni marcate. Quello che poi di fatto è stato ottenuto cattura ben più fenomeni di quelli inizialmente attesi (per un dettaglio, si veda più avanti il capitolo 5.5 relativo ai risultati). Uno strumento come LISCA, l'algoritmo usato per la ricerca, permette infatti di calcolare la probabilità che una certa costruzione sintattica si realizzi nel linguaggio. Grazie a questo approccio è stato possibile ricavare per esempio quali siano le costruzioni marcate, perché saranno quelle che hanno una bassa probabilità di realizzarsi. L'algoritmo inoltre, originariamente progettato per operare su corpora linguistici monolingue, applicato per il confronto di più lingue si è rivelato efficace anche nel far emergere principi universali o di tendenza. I principi universali saranno rappresentati da quei comportamenti che si realizzano in più lingue, mentre i principi di tendenza si possono osservare solo in certi sottogruppi accomunati da una qualche proprietà.

Man mano che nel corso degli anni la comunità scientifica ha avvertito il grande vantaggio che deriva della collaborazione fra linguistica, informatica e statistica, è sorta anche l'esigenza di definire schemi e criteri per l'annotazione morfologica e sintattica di risorse linguistiche che possedessero una valenza universale, che fossero cioè formalmente applicabili a qualsiasi lingua in modo da favorirne il confronto [Nivre, 2015]. Da questa esigenza nacquero alcuni schemi di annotazione, fra cui quello del progetto Universal Dependencies (UD), a cui è dedicato il prossimo capitolo.

Prima di introdurre le metodologie e gli strumenti linguistico-computazionali impiegati in questo studio e prima di parlare del progetto UD, si presenteranno qui le proprietà tipologiche di italiano, inglese e spagnolo, le tre lingue che sono state scelte per svolgere i confronti in questa ricerca.

2.6.1 Le lingue della ricerca

Le lingue su cui è stata effettuata l'indagine oggetto di questo elaborato sono italiano, inglese e spagnolo. Tutte e tre appartengono alla famiglia delle lingue indoeuropee, ma fanno parte di due rami distinti: l'inglese rientra fra le lingue germaniche, mentre italiano e spagnolo sono lingue romanze.

Dal punto di vista tipologico, l'inglese è una lingua SVO flessiva, ma con forti caratteri da lingua isolante che costringono i costituenti a seguire un ordine lineare piuttosto rigido. Un'altra caratteristica delle lingue isolanti è di costruire le parole che rappresentano concetti complessi attraverso la giustapposizione di lessemi che rappresentano concetti più semplici. Il tedesco, altra lingua germanica, è ancora più ricca di nomi composti, ma se ne riscontrano anche un elevato numero di casi in inglese. L'inglese come lingua flessiva è molto peculiare; possiede infatti meno di dieci morfemi flessionali: la *-s* plurale, morfemi per il comparativo e superlativo degli aggettivi, la terza persona singolare del presente, *-ed* del passato e del participio passato, suffisso *-ing* del participio presente e pochi altri.

Italiano e spagnolo appartengono al ramo delle lingue romanze e derivano entrambe dal latino. Si tratta di lingue prevalentemente flessive con ordinamento SVO come l'inglese. Sono lingue che presentano parole internamente complesse dotate però di una base lessicale semplice a cui si appoggiano uno o più affissi flessionali che veicolano valori grammaticali e assommano diverse funzioni. Le lingue come l'italiano e lo spagnolo per loro natura sono piuttosto irregolari e presentano diverse idiosincrasie, quali casi di omonimia, sinonimia, polisemia, etc[8].

Una delle proprietà su cui le tre lingue più divergono è la possibilità di avere o meno omissione del soggetto. La presenza obbligatoria o meno del soggetto all'interno della frase è uno dei criteri su cui si basa la suddivisione fra lingue pro-drop (abbreviazione di *pronoun-dropping*) e non pro-drop. Le lingue pro-drop, anche chiamate "lingue a soggetto nullo", sono quelle in cui alcune categorie di pronomi possono essere omessi quando la frase permette di inferirli pragmaticamente o grammaticalmente; viceversa le lingue non pro-drop prevedono addirittura l'uso di soggetti con funzione espletiva, chiamati generalmente *dummy*, che svolgono solo un ruolo sintattico senza nessun significato semantico. Delle tre lingue della ricerca, italiano e spagnolo sono a possibile soggetto-nullo, mentre l'inglese è una lingua non pro-drop.

[8]Non si tratta di fenomeni totalmente assenti nel caso dell'inglese, ma in italiano e spagnolo si presentano in quantità maggiori.

Capitolo 3

Grammatica a Dipendenze e Universal Dependencies

La rappresentazione delle relazioni di dipendenza sintattica di una risorsa linguistica è un processo che prende il nome di annotazione. Tale operazione è funzionale al miglioramento dei compiti di elaborazione del linguaggio naturale quali estrazione di informazione, *question answering*, o estrazione di dati linguistici. Per questa ragione negli ultimi anni si è assistito ad un crescente interesse verso la definizione di standard per la rappresentazione sintattica a dipendenze, la quale costituisce una strategia semplice e trasparente per la riproduzione della struttura predicato-argomento della frase [Nivre, 2015].

In questo capitolo verrà descritta la teoria che sta alla base della rappresentazione a dipendenze e di come questa sia funzionale all'analisi automatica dei testi. Infine si parlerà del progetto Universal Dependencies che ha sviluppato un *tagset* universale per la rappresentazione della sintassi e che è stato utilizzato in questa ricerca.

3.1 La grammatica a dipendenze

La grammatica a dipendenze è una rappresentazione della sintassi resa popolare in Europa da Tesnière. Si oppone storicamente alla grammatica a costituenti, molto usata invece negli Stati Uniti sin dagli anni '30 grazie a Bloomfield prima e Chomsky poi [Mel'cuk, 1988].

A differenza della rappresentazione a costituenti, la grammatica a dipendenze produce una schematizzazione più adatta ai task di analisi automatica del linguaggio. Si tratta infatti di un formalismo, piuttosto che di una teoria, che permette di rappresentare la frase come un insieme di elementi lessicali collegati fra loro da relazioni binarie asimmetriche, chiamate appunto dipendenze. A differenza della rappresentazione a costituenti infatti, che prevede che gruppi di parole possano comportarsi come elemento unitario, nella rappresentazione a dipendenze sono le singole parole a svolgere funzioni grammaticali nella frase, entrando in relazione con gli altri elementi. Inoltre, nella rappresentazione a dipendenze, manca completamente il nodo-frase, sostituito dal nodo-radice.

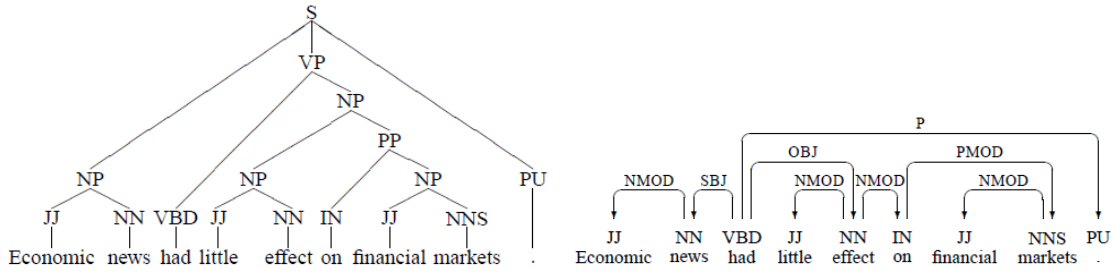


Figura 3.1: Rappresentazione a costituenti (sinistra) e a dipendenze (destra) di una frase della Penn Treebank.

Per comprendere appieno la nozione di dipendenza è utile fare riferimento alla descrizione di Tesnière che si trova nei capitoli iniziali di "*Elementi di Sintassi strutturale*" [1959]:

"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]."

Le relazioni di cui parla Tesnière si creano fra una testa (anche chiamata *governor* o reggente) e un dipendente (altrimenti detto modificatore) e si basano su criteri, più o meno vincolanti in base al tipo di relazione, che rappresentano le modalità in cui testa e dipendente interagiscono.

Tesnière fa riferimento nella sua definizione a processi mentali innati che permettono all'individuo di interpretare e comprendere la semantica della frase basandosi sul modo in cui ciascun elemento entra in relazione con gli altri. Se invece intendiamo la grammatica a dipendenze come un formalismo usato per la rappresentazione esplicita della sintassi, è chiaro che le relazioni che collegano teste e dipendenti non sono più semplicemente innate, ma devono essere definite per mezzo di criteri tali da fornire indicazioni non ambigue per il loro trattamento. La definizione di quali etichette usare per la rappresentazione delle relazioni grammaticali e le indicazioni su come applicarle costituisce uno schema di annotazione. L'applicazione di uno schema di annotazione ad una risorsa linguistica permette di intercettare le informazioni linguistiche che in essa sono contenute.

Il modo in cui vengono definiti i principi combinatori degli elementi è di centrale importanza dal momento che è sulla base di ciò che si stabiliscono le relazioni di dipendenza. Il rischio che comporta non possedere delle linee guida non ambigue e contraddittorie è quello di far sì che vengano assegnate annotazioni diverse a costruzioni invece equivalenti. Sarebbe invece auspicabile, nel momento in cui si vuole poter sfruttare il formalismo della grammatica in una prospettiva universale o di studio tipologico, che le disomogeneità di rappresentazione vengano il più possibile evitate, sia fra lingue che internamente alla stessa lingua.

Tra gli schemi di annotazione che sfruttano il formalismo della grammatica a dipendenze, le principali dimensioni di variazione che si possono individuare riguardano la scelta delle categorie grammaticali e la conseguente definizione di diversi criteri di annotazione per quelle costruzioni per cui non si individuano teorie linguistiche di riferimento, la selezione della testa a livello semantico o sintattico e l'ammissione o meno del vincolo di *proiettività* [Nivre, 2005]. Si tratta in ciascun caso di criteri fondamentali che definiscono i principi base dello schema di annotazione e che si

fondano, negli ultimi due casi, su nozioni linguistiche. Ciascuno dei tre punti sarà ora affrontato nel dettaglio.

La selezione di quali fenomeni linguistici intercettare per mezzo delle relazioni di dipendenza e, prima ancora, dell'assegnamento delle categorie grammaticali è forse la più significativa delle variazioni che possono verificarsi fra i diversi schemi di annotazione. A questo è poi legata anche la definizione dei criteri per l'applicazione di tali etichette, di cui il trattamento della coordinazione costituisce un caso particolare poiché non esiste nessuna teoria linguistica che guidi al trattamento di tale costruzione. Le congiunzioni e le strutture coordinate possono risultare difficili da trattare attraverso il rigido formalismo della rappresentazione a dipendenze e quindi esistono diverse proposte per risolverne le ambiguità: la coordinazione, per definizione, pone due elementi sullo stesso livello sintattico, ed è quindi difficile stabilire univocamente quale sia il candidato migliore per svolgere il ruolo di testa della relazione. La scelta può ricadere sul primo elemento della sequenza coordinata, sull'ultimo o sulla congiunzione stessa. Ciascuna delle tre rappresentazioni può essere interpretata come corretta, a patto che sia coerente in tutti i casi di congiunzione a cui viene applicato lo stesso schema. La coerenza è infatti una delle caratteristiche fondamentali di uno schema di annotazione: l'inconsistenza di trattamento non permette di rintracciare all'interno di una risorsa linguistica annotata le costruzioni equivalenti.

I principi per la selezione della testa di una relazione si basano solitamente sia su criteri semantici che sintattici, ma non sempre i due producono un risultato univoco. La selezione della testa è di fondamentale importanza, difatti la stessa frase, con l'applicazione di principi diversi per la selezione della testa, può risultare in due analisi molto lontane. Di seguito si riportano i criteri per l'identificazione di una relazione sintattica tra una testa T e un dipendente D in una costruzione C proposti da Zwicky [1985] e Hudson [1984].

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;

6. La posizione lineare di D è specificata in riferimento a T.

Uno schema di annotazione che vuole proporsi come standard universale deve preferibilmente scegliere di assegnare il ruolo di testa sintattica alle parole piene in modo da massimizzare i parallelismi possibili fra lingue. Difatti, se si scegliesse di adottare le categorie grammaticali funzionali come teste delle relazioni, il confronto fra lingue analitiche e sintetiche non sarebbe possibile e, per queste ultime, bisognerebbe definire comunque dei criteri di annotazione differenti.

Infine, per spiegare il vincolo di proiettività, è indispensabile prima presentare il concetto di albero sintattico. Sebbene le teorie sulla grammatica a dipendenze differiscano sotto molti aspetti, tutte concordano sul fatto che una frase sia costituita da una serie di nodi lessicali e la sintassi sia rappresentata dalla struttura gerarchica descritta dagli archi che collegano tali nodi [Carnie, 2013]. Le relazioni contenute in ogni singola frase costituiscono dunque un grafo aciclico orientato, ovvero una struttura ad albero radicato, che permette di rendere espliciti i rapporti di gerarchia fra elementi grammaticali, solitamente non corrispondenti all'ordine lineare che essi assumono nell'enunciato.

Una delle possibili conseguenze di ricreare la gerarchia di relazioni attraverso un albero sintattico è quella di incappare nel vincolo di proiettività [Lecerf, 1960, Hays, 1964, Marcus, 1965]. Questo vincolo viene soddisfatto dall'albero quando ciascuna parola che nell'ordine lineare si trova fra una testa T e il suo dipendente è comunque dominata da T, ovvero fa parte del sotto-albero che ha T come radice. Le grammatiche a dipendenze tendono a soddisfare questo vincolo, dal momento che non rispettarlo compromette l'efficienza dei parser [Nivre and Nilsson, 2005], sebbene a volte risulti troppo rigido per rappresentare lingue con ordine delle parole libero o molto flessibile [Nivre, 2005].

Indipendentemente dai tratti di variabilità individuale, ciascuno schema di annotazione può essere visto come il risultato della combinazione di informazioni linguistiche di base trasversali ai vari livelli di analisi linguistica. Gli schemi di annotazione sono quindi un insieme di informazione categoriale, strutturale e relazionale. L'informazione categoriale deriva dall'assegnamento di categorie alle unità e relazioni linguistiche identificate in un testo (e.g. verbo, sintagma nominale, soggetto, etc.). L'informazione strutturale si ricava invece dall'identificazione di strutture che possono essere sia interne ad un singolo token (e.g. le sue proprietà morfologiche), sia

raggrupparne più insieme (e.g. i costituenti sintagmatici dell'annotazione sintattica). L'informazione relazionale infine è relativa alla definizione di relazioni tra le unità linguistiche identificate, ad esempio le relazioni di dipendenza sintattica (e.g. soggetto, oggetto diretto).

Ad oggi esistono diversi formalismi che si basano sulla teoria della grammatica a dipendenze, ognuno basato su un principio diverso: fra questi i più noti sono Word Grammar (WG) [Hudson, 1984], Functional Generative Description (FGD) [Sgall, 1967], Dependency Unification Grammar (DUG) [Hellwig, 1986]. La teoria della WG assegna un ruolo centrale alle singole parole, che contengono quasi totalmente l'informazione sulla struttura della frase; la sintassi di conseguenza non è altro che una serie di principi che derivano dal combinarsi delle parole. La FGD invece interpreta le frasi come composte di molteplici livelli linguistici che si intrecciano per comporre la sintassi. La DUG infine è una versione della rappresentazione a dipendenze dove ciascun arco è arricchito di informazioni che ne dovrebbero facilitare l'elaborazione computazionale. Ciascuna di queste tradizioni teoriche è accomunata dal principio che la parte essenziale della struttura sintattica della frase risieda in relazioni binarie asimmetriche tra elementi lessicali [Nivre, 2005].

3.2 Annotazione manuale e parsing a dipendenze

Il rigido, e allo stesso tempo semplice, formalismo della grammatica a dipendenze e la rappresentazione mediante alberi sintattici stanno riscuotendo molta fortuna nell'ambito del trattamento automatico del linguaggio [Kubler et al., 2009]. Questo di conseguenza ha comportato anche un grande interesse verso lo sviluppo di sistemi computazionali in grado di predire e riprodurre tali formalismi, facilitato anche dal costante aumento delle risorse linguistiche annotate manualmente a cui si è assistito negli ultimi anni [Buchholz and Marsi, 2006, Nivre et al., 2007b].

Gli strumenti che permettono di realizzare un'annotazione linguistica in maniera automatica sono i *parser*. Il processo di parsing consiste appunto nel definire, dato un flusso continuo di testo in input, la struttura sintattica delle singole frasi e le relazioni fra le parole[1] che le compongono. La grammatica a dipendenze costituisce una delle possibili rappresentazioni sintattiche che un parser può realizzare. Tuttavia

[1]Quando si tratta di parsing, è più corretto parlare di *tokens* piuttosto che di parole.

in questo caso il formalismo della teoria linguistica viene ulteriormente semplificato; per questa ragione è più corretto parlare di rappresentazione a dipendenze, piuttosto che di parsing con grammatica a dipendenze.

Non si entrerà qui nel dettaglio degli algoritmi che vengono usati per svolgere il parsing e degli strumenti a disposizione, ma si cercherà di presentare i principi generali secondo cui operano e, nel prossimo capitolo, le maggiori difficoltà legate a questo task [2].

Il parsing delle relazioni sintattiche non è mai fine a se stesso, ma è funzionale alla realizzazione di compiti successivi, come estrazione di informazione, traduzione automatica e *question answering*, che comportano anche l'analisi della componente semantica del testo. Per questa ragione individuare le regolarità nell'uso delle strutture sintattiche e renderle esplicite in modo tale che diventino sfruttabili da strumenti per l'analisi automatica è fondamentale. Importante quindi avere a disposizione uno strumento di parsing utile ed efficace, riconoscibile dal soddisfacimento di questi quattro requisiti [Nivre, 2015]:

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.

Accanto a questi requisiti si affiancano anche dei parametri quantitativi per la valutazione dell'efficienza di un parser. La metrica di valutazione delle performance prende in considerazione i valori di LAS (*label attachment score*), UAS (*unlabeled attachment score*) e LA (*label attachment*) [Chen et al., 2014]. Rispettivamente questi indici si riferiscono a [Montemagni, 2013]:

- La proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda sia la testa sintattica sia il ruolo (tipo di relazione di dipendenza) svolto in relazione ad essa (LAS);

[2] Per una rassegna più dettagliata si veda, fra gli altri, Nivre [2005].

- La proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione della testa sintattica.(UAS);
- La proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione del tipo di dipendenza (LA).

Per valutare l'accuratezza nel riconoscimento delle singole classi delle relazioni di dipendenza si utilizzano i valori di *precision* e *recall*. La *precision* è calcolata come il rapporto fra il numero di elementi etichettati correttamente e il numero totale di archi a cui è stata assegnata quella data etichetta. La *recall* invece è il rapporto fra il numero di elementi etichettati correttamente e il numero totale di elementi del test set che effettivamente presentano quella relazione (indipendentemente dalla valutazione del parser). Solitamente si fa riferimento anche al valore di *F-score*, ovvero la media armonica fra *precision* e *recall*.

Per quanto riguarda l'approccio usato dai parser, Carroll [2000] individua due possibili paradigmi che stanno alla base dei sistema di analisi sintattica automatica: uno si basa sull'applicazione di regole formali che costituiscono un'approssimazione del linguaggio naturale (approccio *grammar-driven*), il secondo invece estrae le regole attraverso analisi statistiche indotte dai dati veri e propri (approccio *data-driven*). Mentre nei *grammar-driven parser* ogni frase da analizzare deve far parte del linguaggio definito dalla grammatica, i sistemi *data-driven* sono capaci di indurre i vincoli per guidare il processo di analisi delle frasi senza creare un linguaggio entro il quale restringere le possibili frasi da analizzare. In questo secondo gruppo di parser, due sono le tecniche più comuni che vengono usate dagli algoritmi di parsing. Il primo è un sistema di analisi *shift-reduce* basato sulle transizioni *(shift-reduce transition-based parsing)* in cui gli elementi della frase vengono analizzati seguendo l'ordine lineare di apparizione da sinistra a destra [Nivre et al., 2007b]. Nel secondo metodo invece la frase viene processata come insieme e l'analisi assegnata ad essa corrisponde all'albero di dipendenza che ha ottenuto il più alto valore di probabilità. In questo secondo metodo il processo è normalmente più lungo perché per ogni frase vengono generati ed esplorati tutti i possibili alberi sintattici [Bohnet, 2010]. L'analisi corretta per una frase viene inferita a partire da un training corpus di riferimento da cui estrarre le statistiche che permettono allo strumento di risolvere il problema della disambiguazione attraverso l'assegnamento di un punteggio di probabilità di analisi ad ogni possibile interpretazione della frase. Risulta indispensabile per lo sviluppo di questi strumenti la possibilità di avere a disposizione delle risorse linguistiche dal-

le quali tali parser possano "apprendere" quale informazione linguistica associare al testo [Dell'Orletta et al., 2011].

I parser per l'annotazione automatica dei testi non potrebbero raggiungere alti punteggi di accuratezza se non avessero a disposizione dei modelli corretti su cui essere addestrati. Tali modelli sono realizzati attraverso l'annotazione morfo-sintattica manuale di corpora testuali. L'annotazione linguistica manuale dei testi, sebbene costosa in termini di tempo e risorse, mette a disposizione dei dati che possono essere assunti come *gold*, ovvero privi di errori, da usare in fase di addestramento.

Il processo di annotazione può essere realizzato a partire da testi grezzi (vedi sottosezione 3.1.1) oppure può consistere nella revisione dell'analisi svolta da un parser automatico. Solitamente l'annotazione o la revisione di un testo viene svolta indipendentemente da due o più annotatori, i quali poi si confrontano sul risultato della loro attività. Per avere un dato oggettivo, viene calcolato l'indice di accordo fra i due soggetti (*Interannotator Agreement*) [Artstein and Poesio, 2008], ovvero un coefficiente che quantifica la concordanza degli annotatori nell'assegnazione delle etichette per una determinata categoria. Manning [2011] osserva che, sebbene in letteratura si dica che il valore di disaccordo fra annotatori si aggira mediamente intorno al 3%, questo indice sale a 7% se il task non prevede la revisione di un testo pre-annotato bensì l'annotazione *ex novo* di testi tokenizzati. Un certo grado di disaccordo pare essere fisiologico, dovuto ad ambiguità del testo o a cali di attenzione da parte dei soggetti che svolgono il task. A questo però si aggiunge anche il ben noto problema dell'*ancoraggio* ovvero un fenomeno dovuto ad un pregiudizio mentale per cui i soggetti sovrastimano la performance di un tagger o di un parser e tendono a non modificare l'annotazione esistente [Berzak et al., 2016][Tsarfaty et al., 2011].

Al momento le treebank gold sono disponibili in un numero ristretto di lingue, mentre è in crescita l'interesse per lo sviluppo di strumenti multilingua che possano quindi operare correttamente indipendentemente dalle variazioni fra lingue [Buchholz and Marsi, 2006]. La mancanza di criteri universali per l'annotazione delle lingue e la conseguente assenza di strumenti che possano operare su un piano multi-lingua creano dei grandi ostacoli per il progresso della disciplina. Per esempio è impossibile confrontare risultati empirici fra lingue diverse o sostenere lo sviluppo di un parser universale se manca uno standard per l'annotazione. Questo fatto compromette anche l'avanzare di studi in una prospettiva tipologica, che sembrano invece essere un campo di indagine promettente [Petrov et al., 2011].

Dalla consapevolezza di queste urgenze e problematiche è nato il progetto delle Universal Depenedencies: esso rappresenta una delle ultime iniziative di annotazione linguistica avviate a livello internazionale per la definizione di dati e schemi di annotazione da poter utilizzare come standard di riferimento multilingue.

3.3 Il progetto Universal Dependencies

Il progetto Universal Dependencies[3] (UD) nasce nel 2014. La prima release è del 2015 e comprendeva allora 10 lingue; il 1 marzo 2017 è stata rilasciata la versione 2.0 che attualmente contiene 70 treebank e 50 lingue diverse. Per questa ricerca ci si è serviti della versione 1.2, rilasciata a novembre 2015 e composta da 37 treebank che rappresentano 33 lingue. Il progetto fu avviato al fine di creare un valido modello di annotazione grammaticale inter-linguistica (su un doppio livello morfologico e sintattico) e di fornirne le linee guida per una sua applicazione.

UD rappresenta un'evoluzione dei precedenti modelli di annotazione morfo-sintattica universale: lo schema infatti eredita le relazioni di dipendenza dall'iniziativa Universal Stanford Dependencies [De Marneffe et al., 2006], le etichette per le categorie grammaticali dal Google Universal Part-Of-Speech Tagset [Petrov et al., 2011] e le feature morfologiche da Interset [Zeman, 2008]. L'obiettivo fondamentale del progetto è di affermarsi come standard *de facto* sostituendo tutti gli schemi fin'ora esistenti e, per farlo, si propone di fornire un tagset che permetta di enfatizzare le similarità fra lingue senza però appiattirne le differenze quando necessario [Nivre et al., 2016]. Questo comporta ovviamente di dover giungere a dei compromessi con le teorie linguistiche, di cui lo schema UD non vuole diventare un sostituto. La rappresentazione che offre è funzionale alla schematizzazione della sintassi, la quale, così rappresentata, può però scontrarsi con alcuni *framework* di grammatica teorica per quanto riguarda le strutture profonde della frase [Nivre, 2015]. Tali obiettivi teorici ben si accordano con la maggior parte dei principi fondanti del progetto, riportati qui di seguito.

1. UD deve soddisfare le esigenze di analisi linguistica per le singole lingue;
2. UD deve fornire una base linguistica che permetta di far emergere i parallelismi fra le lingue e le similarità fra tipi linguistici;

[3] http://universaldependencies.org/ (consultato il 10/04/2017).

3. UD deve consentire una rapida e consistente annotazione da parte degli annotatori (umani);

4. UD deve consentire di realizzare il parsing automatico della frase con alti valori di accuratezza;

5. UD deve essere facilmente comprensibile e utilizzabile da chiunque, a prescindere dal background;

6. UD deve essere di supporto ai successivi task di elaborazione del testo.

3.3.1 UD: principi di annotazione

L'annotazione in UD si basa sul formalismo della rappresentazione sintattica delle relazioni di dipendenza, di cui si sono già visti i vantaggi nell'ambito dell'elaborazione automatica dei testi. Allo stesso tempo segue anche un principio lessicalista, per cui le parole sono le unità minime dell'analisi grammaticale. Le parole sono dotate di proprietà morfologiche e sono legate da relazioni di dipendenza sintattica: le connessioni fra parole sono ciò che lo schema di annotazione UD cerca di rendere esplicito [Nivre et al., 2016].

Le risorse linguistiche messe a disposizione da UD sono costruite attraverso il progressivo susseguirsi di certe operazioni. Queste sono, in ordine, tokenizzazione, annotazione morfologica e annotazione sintattica.

Tokenizzazione

Le parole che svolgono una funzione sintattica non sempre coincidono con le unità ortografiche separate da spazi bianchi. Seguendo il principio lessicalista, gli elementi fondamentali su cui si basa l'annotazione UD sono le parole sintattiche, ovvero le unità minime di analisi linguistica, anche chiamate tokens [Lenci et al., 2005]. La tokenizzazione, ovvero la segmentazione del testo in tokens, rappresenta dunque il primo passaggio per poter realizzare l'analisi della frase. Si tratta di un processo fortemente vincolato dalla tipologia e dalla distribuzione specifica della lingua, quindi 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).

Annotazione morfologica

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

- un *lemma*, corrispondente al contenuto semantico della parola e determinato da dizionari e lessici specifici.
- 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 sono generalmente determinati da dizionari e lessici specifici per ogni lingua, mentre le *part-of-speech tag* e le proprietà grammaticali vengono estratti da un inventario universale, predefinito. Le *feature* forniscono informazioni aggiuntive riguardo la parola, la sua categoria grammaticale e proprietà morfosintattiche. Ciascuna *feature* si presenta nella forma Nome=Valore e non ci sono restrizioni sul numero di tratti linguistici che si possono applicare ad una parola. Di seguito si riportano le tabelle con la lista di POS e *feature* disponibili nella versione 1.2. Nella versione 2.0 tuttavia non sono state fatte modifiche significative a queste due categorie.

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	

Figura 3.2: Part-of-speech tag universali.

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

Figura 3.3: Tratti linguistici universali (*features*).

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. Le relazioni previste dal tagset di UD hanno lo scopo di catturare il maggior numero di funzioni grammaticali trasversali al maggior numero di lingue, massimizzando i parallelismi di tipo sintattico.

Per consentire un confronto multilingua efficace, 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 [Nivre, 2015]. Questo significa che solo le teste lessicali sono deducibili dalla struttura ad albero, mentre gli elementi funzionali sono rappresentati come sottocategorie delle parole piene. Ciò consente di avere annotazioni comparabili per le stesse relazioni grammaticali che, di lingua in lingua, possono essere espresse per mezzo di classi funzionali, elementi morfologici o non essere marcate affatto [Nivre, 2016]. Tuttavia UD mette comunque a disposizione delle relazioni per catturare fenomeni specifici delle lingue, rappresentate come relazioni sottospecificate di quelle universali.

Le 40[4] diverse relazioni sintattiche sono classificate secondo un principio di differenziazione strutturale che distingue frasi di tipo nominale da frasi di tipo predicativo; in una categoria distinta ricadono gli altri tipi di modificatori. Lo schema è riportato nella tabella in Figura 3.4, già descritta in Nivre [2015]. Le relazioni possono avere una testa verbale o nominale, oppure non avere una testa come nel caso della relazione speciale *root*. Fra le possibili dipendenze con testa verbale, alcune relazioni sono *core* (fascia alta), *non-core* (fascia mediana) e speciali (fascia bassa). Le relazioni marcate con asterisco (*) hanno come dipendente un verbo ausiliare. Si noti inoltre che viene fatta una distinzione fra diatesi attiva e passiva per quanto riguarda la relazione soggetto e i verbi ausiliari.

	Nominal Dep	**Predicate Dep**	**Other Dep**
Predicate Head	nsubj nsubjpass dobj iobj	csubj csubjpass ccomp xcomp	
	nmod	advcl	advmod neg
	vocative discourse expl dislocated	aux* auxpass* cop*	mark punct
Nominal Head	nummod appos nmod	acl	amod det neg case
No Head		root	

Compounding	**Coordination**	**Other**
compound name mwe goeswith	conj cc punct	list parataxis remnant reparandum foreign dep

Figura 3.4: Relazioni di dipendenza sintattica (UD v1.2).

Formato CoNLL

Il progetto UD mette a disposizione le risorse linguistiche annotate in formato CoNLL-X codificate in file *plain text* (UTF-8). Si tratta quindi di file *gold*, ovvero idealmente privi di errori e ottenuti grazie al lavoro di annotatori umani, che hanno potuto quindi svolgere l'annotazione facendo affidamento sulle loro competenze di parlanti nativi e sulle conoscenze extra-linguistiche.

[4]Nella versione 1.2.

Nel file, le righe vuote indicano i confini di parola, le righe di commento sono introdotte dal simbolo #, mentre le righe di parola contengono l'annotazione di un token, al quale sono associate diverse tipologie di informazione contenuta ognuna in un campo separato dagli altri attraverso una singola tabulazione. I campi rappresentano le seguenti informazioni:

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.

3.3.2 Le treebank della ricerca

Per questo studio sono state utilizzate le tre treebank descritte qui di seguito.

English Web Treebank La treebank inglese è costituita da 16,624 frasi e 254,830 tokens e contiene testi estratti dal Web: post di blog, email, Q&A, e recensioni. La risorsa è stata inizialmente creata dal Linguistic Data Consortium (LDC) rispettando lo standard Stanford Dependencies e Google part-of-speech tagset. Successivamente è stata convertita nello standard UD [Silveira et al., 2014].

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*, di conversione e di armonizzazione che hanno portato alla IUDT (acronimo per *Italian Universal Dependency Treebank*). 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. La TUT *(Turin University Treebank)* [Bosco et al., 2000], e la ISST-TANL, inizialmente rilasciata come la ISST-CoNLL per la *CoNLL-2007 shared task* [Montemagni and Simi, 2007], sono state le principali risorse per la costruzione della MIDT (*Merged Italian Dependency Treebank*) [Bosco et al., 2012], la quale è stata successivamente convertita nella ISDT (*Italian Stanford Dependency Treebank*) [Bosco et al., 2013]. La ISDT nacque per essere la prima risorsa italiana annotata secondo il formalismo delle *Stanford Dependencie* e fu rilasciata in occasione del *dependency parsing shared task di Evalita 2014*. Venne poi utilizzata come punto di partenza per la definizione, tramite conversione, di IUDT, il corpus annotato secondo il modello delle Universal Dependencies, pubblicato per la prima volta nel gennaio 2015 (come dichiarato sul sito ufficiale UD).

La risorsa è costituita da 13,815 frasi che si suddividono in 325,816 tokens estratte da testi appartenenti a diversi generi testuali: articoli giornalistici, testi di ambito giuridico e articoli di giornale scritti in italiano semplificato.

Spanish UD treebank La risorsa della lingua spagnola è il prodotto della conversione automatica nello standard UD dell'Universal Google Dataset (versione 2.0). Conta di circa 4,000 frasi e 112,718 tokens estratti da articoli di quotidiani e blog [McDonald et al., 2013].

Capitolo 4

Strumenti e metodologie

Dopo aver trattato dell'annotazione manuale, del parsing sintattico e del progetto che mette a disposizione le risorse annotate e le linee guida per svolgere tali task, in questo capitolo verranno presentate non solo alcune delle difficoltà legate alla presenza di errori all'interno delle risorse e dei risultati del parsing automatico, ma anche gli strumenti che sono stati messi a punto per fronteggiare tali problemi. In particolare si parlerà di LISCA, un algoritmo di ordinamento degli archi sulla base della plausibilità e che è stato adoperato per questa ricerca.

4.1 Errori di annotazione

Le performance di annotazione dei parser sintattici, seppur molto alte e soddisfacenti, non raggiungono mai la perfezione. Solitamente l'accuratezza di un sistema di analisi morfo-sintattico che raggiunge lo stato dell'arte si aggira intorno al 97-98% per quanto riguarda l'assegnazione delle classi grammaticali, ovvero il 2-3% delle valutazioni grammaticali che assegna sono sbagliate[Manning, 2011]. Per quanto riguarda il parsing sintattico invece i valori di accuratezza sono inferiori e legati al tipo linguistico: Nivre et al. [2007a] riportano valori di LAS (vedi capitolo 3.2) compresi fra 84% e 90% per catalano, cinese, inglese e italiano e fra 76% e 80% per arabo, basco, ceco, greco, ungherese e turco[1]. Un certo margine di errore sembra essere fisiologico: spesso la lingua presenta ambiguità difficili da risolvere anche per un essere umano e, qualora fossero invece risolvibili, spesso per farlo il parlante si serve

[1]I dati sono confermati anche dal più recente de Lhoneux et al. [2016].

del contesto extra-linguistico, su cui la macchina non può fare affidamento. In altri casi tuttavia l'errore può essere dovuto solo ad una errata inferenza del parser, fatte sulla base del modello statistico usato.

Il valore di accuratezza peggiora ulteriormente quando un sistema statistico di parsing addestrato su un certo modello della lingua (i.e. un genere o un registro) viene applicato ad un testo il cui dominio è molto diverso da quello del modello di addestramento [Gildea, 2001]. L'annotazione manuale di corpora specifici per il dominio di interesse è sicuramente una tattica efficace per risolvere tali problemi, ma anche eccessivamente costosa, per questo raramente adottata [Dell'Orletta et al., 2011]. Inoltre non è neppure chiaro cosa si intenda per dominio e quali elementi di un testo lo definiscano precisamente. Il dominio può anche non essere un genere testuale o un particolare registro, bensì una lingua vera e propria, e l'obiettivo in questo caso è quello di utilizzare uno strumento addestrato su una certa lingua per svolgere operazioni su un altra. La disciplina che si occupa di trovare tecniche per la risoluzione di tali questioni prende il nome di *Domain Adaptation* (DA) [Plank, 2016].

Da un punto di vista più generale e meno legato ai problemi di DA , essendo il parsing un'operazione che si svolge in più fasi consecutive, un errore nella prima fase comporta una serie di sbagli che ricadono a cascata sui livelli successivi; infine tali errori di annotazione influiscono direttamente sui compiti di estrazione dell'informazione dai testi, successivi al parsing, con un forte impatto negativo [Boyd et al., 2008]. Per questa ragione la comunità scientifica si dedicata attivamente alla risoluzione del problema nel tentativo di trovare metodi automatici di individuazione degli errori o per ridurne il numero. Per esempio Sagae and Tsujii [2007] propongono di confrontare il risultato di più parser e selezionare come analisi per la frase quella condivisa dal maggior numero di essi. Questo può ridurre il numero di errori totali, ma non lo azzera del tutto.

Nei testi annotati con una rappresentazione delle relazioni sintattiche attraverso archi di dipendenza, per errori si intendono quegli archi a cui è stata assegnata un tipo di relazione sbagliata, oppure quelli che creano una relazione fra due elementi che non sono realmente in relazione di dipendenza l'uno con l'altro. Gli errori possono anche essere commessi a livello di separazione del testo in parole sintattiche (per esempio in italiano sono i pronomi clitici a rappresentare le maggiori difficoltà per i tokenizzatori) così come in fase di assegnazione delle categorie grammaticali. Addirittura si osserva che la maggior parte degli errori commessi dal parser sintattico sono in realtà causati da interpretazioni scorrette in uno dei livelli precedenti.

In letteratura si possono rintracciare tre diversi approcci per il rilevamento degli archi sbagliati: basati su regole, supervisionati e non supervisionati. Il primo metodo prevede la comparazione di regole della grammatica *gold* con le regole dedotte dai dati automaticamente analizzati [Dickinson, 2010, Dickinson and Smith, 2011, Ambati et al., 2010]. L'approccio supervisionato invece si serve unicamente di regole gold. Tra gli esempi, Kawahara and Uchimoto [2008] si servono di un classificatore esterno per valutare la plausibilità di ciascuna frase analizzata, mentre Dickinson and Meurers [2003] ottengono statistiche da una grammatica gold ed estraggono quelle frasi che più si discostano dalle regole dedotte. Altri strumenti ancora sviluppano dei classificatori di errori addestrati su alberi sbagliati generati automaticamente che fungono da metro di confronto per quelli corretti della treebank [Attardi and Ciaramita, 2007, Anguiano and Candito, 2011, Hall and Novak, 2005]. L'approccio non supervisionato invece si serve di statistiche ricavate automaticamente da grandi quantità di dati che permettano all'algoritmo di assegnare un punteggio di affidabilità a ciascun fenomeno linguistico [Dell'Orletta and Venturi, 2016]. Solitamente si tratta di indici calcolati per mezzo della *Local Mutual Information* (o una funzione ad essa vicina) relativa a specifiche configurazioni sintattico-lessicali con finalità di auto-apprendimento [Van Noord, 2007]. Quest'ultimo è l'approccio adottato da LISCA [Dell'Orletta et al., 2013].

Gli algoritmi che assegnano dei punteggi di plausibilità permettono, fra le altre cose, di semplificare l'operazione di revisione automatica dell'annotazione: assumendo che, sebbene costosa, la revisione manuale sia il metodo di correzione delle risorse annotate più efficace, l'ausilio di questi strumenti permette di estrarre in maniera automatica dai testi analizzati solo quelle costruzioni che effettivamente richiedono una revisione perché anomale, escludendo le più semplici e banali che sarebbero solo uno spreco di tempo ed energie [Dickinson, 2010].

4.2 LISCA

LISCA (*LInguistically–driven Selection of Correct Arcs*) è uno strumento di valutazione di tipo *data-driven* che viene applicato nell'ambito del parsing a dipendenze al fine di migliorare le performance complessive assegnando un punteggio di plausibilità a ciascun arco prodotto dal parser. Lisca è stato sviluppato presso l'ItaliaNLP lab, interno all'istituto di Linguistica Computazionale del CNR di Pisa, ed è stato testato

con successo su due dataset appartenenti a domini distinti; in tutti gli esperimenti ha superato per prestazioni diversi modelli standard, dimostrando così di essere in grado non solo di rilevare archi corretti, ma anche di intercettare peculiarità di dominio. L'approccio data-driven su cui si basa LISCA si oppone tradizionalmente all'approccio *rule-based*: il primo metodo viene usato da strumenti che estraggono delle regole a partire da treebank, ovvero esempi di testi annotati; l'approccio basato su regole invece prevede di verificare che i testi rispettino le norme di una certa grammatica fornita a priori [Nivre et al., 2007b]. Nel caso di LISCA quindi non è necessario creare manualmente grammatiche o risorse per permetterne il funzionamento.

A differenza degli altri approcci non supervisionati che operano sui dati analizzati sintatticamente, lo scopo di LISCA è di ordinare in modo decrescente gli archi riconosciuti in fase di analisi per plausibilità senza far riferimento all'efficacia che una certa costruzione può avere nella risoluzione di un task specifico come il miglioramento di una risorsa annotata. LISCA realizza la sua valutazione sulla base di proprietà linguistiche desunte da un ampio corpus: si tratta di una combinazione di tratti strutturali linguisticamente motivati e vere e proprie proprietà linguistiche catturate a partire dall'albero sintattico della frase e usate per calcolare un punteggio di qualità per ciascun arco.

Lo scopo originale della creazione di un tale strumento riguardava non solo il miglioramento dell'accuratezza dei parser statistici identificando le costruzioni più critiche fra quelle restituite, ma anche di adattare i parser a corpora appartenenti a domini diversi rispetto a quelli di addestramento. Inoltre ovviamente una possibile applicazione riguarda il supporto nella creazione di treebank manualmente annotate: isolare le aree problematiche per il pre- e post-processing umano permette di focalizzare l'attenzione sulle costruzioni che necessitano di essere corrette. Con il tempo tuttavia sono emersi altri possibili utilizzi di questo strumento: la ricerca degli errori è solo una delle possibilità che offre e che è stata sfruttata nella ricerca descritta in questo elaborato.

L'ordinamento di tutti gli archi di una treebank per plausibilità permette di immaginare che, in fondo alla classifica, non si trovino solo gli archi sbagliati, ma anche le costruzioni sintattiche anomale e poco frequenti che sono in qualche maniera deviazione dalla canonicità della lingua. Se applicato su un corpus gold, ovvero privo di errori, l'algoritmo dovrebbe restituire una classifica di archi, tutti corretti, ordinati sulla base della probabilità che si realizzino nella lingua: i punteggi più bassi andranno ad intercettare quei casi in cui, per scopi comunicativi o enfatici, la frase

non rispetta lo standard del proprio registro o della grammatica. Non si tratta in ogni caso di frasi non grammaticali, ma piuttosto frasi che per una qualche ragione hanno delle peculiarità. Tuttavia bisogna tener presente che nessun corpus è mai completamente privo di errori: si è già parlato del fatto che anche le risorse annotate o revisionate manualmente in realtà presentano sempre un certo margine di disaccordo fra coloro che vi lavorano.

È importante sottolineare che per plausibilità di un arco non si intende semplicemente un valore calcolato sulla base della frequenza di occorrenza nel corpus, che ovviamente costituisce comunque un indice con cui fortemente correla. La plausibilità è un valore ben più complesso, dato dal prodotto di diversi fattori, linguistici e strutturali, che co-occorrono e possono rendere un arco di dipendenza più o meno prevedibile in un dato contesto (si approfondiranno nelle prossime sezioni quali siano per LISCA questi fattori). Per questa ragione può accadere che un arco che rappresenta una relazione di fatto corretta, a causa di un errore commesso su un altro elemento o a causa della struttura inusuale della frase, ottenga un punteggio di plausibilità basso pari a quello di archi sbagliati.

Una dimostrazione viene fornita dai seguenti esempi, di cui si mostra la rappresentazione sintattica a dipendenze. La visualizzazione è ottenuta grazie a DG Annotator[2], che permette di visualizzare gli archi di dipendenza e le classi grammaticali assegnati ad ogni token.

La prima frase ("Due le persone sorprese dalla polizia mentre erano intente a giocare con apparecchiature che consentivano vincite da diverse migliaia di lire") è chiaramente una frase estrapolata da un articolo di giornale. L'ambito giornalistico è certamente rappresentativo del linguaggio standard, ma è anche ricco di elisioni e costrutti nominali, tipiche proprio di questo genere. In questa frase appunto il verbo essere, che fungendo da copula avrebbe dovuto accompagnare la radice, è sottinteso ("Sono due le persone [...]"). Questa omissione del verbo comporta per esempio che il soggetto "le persone" sia invece marcato con la relazione *xcomp*, riservata alle subordinate. Questo errore nella frase principale e il fatto che appaia annidato nell'albero sintattico, fanno sì che la relazione dell'oggetto diretto *dobj*, seppur corretta, non venga valutata come plausibile.

Allo stesso modo la seconda frase ("Per le aziende e le società di piccole e medie dimensioni c'è una scelta decisiva:") una relazione semplice come quella di modifica-

[2]http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/ (visitato il 06/04/2017)

zione aggettivale ottiene invece la valutazione di relazione complessa. In questo caso si noti che la punteggiatura è piuttosto anomala: la frase è stata spezzata dopo i due punti e non si conclude quindi con un punto fermo come dovrebbe. Inoltre il soggetto si trova in posizione post-verbale, fatto piuttosto anomalo come si vedrà in seguito (sezione 5.4.1). In posizione pre-verbale si trovano invece due modificatori nominali in sequenza. L'annotazione della frase è corretta ma la struttura complessiva non è standard, di conseguenza LISCA le assegna un punteggio di plausibilità basso.

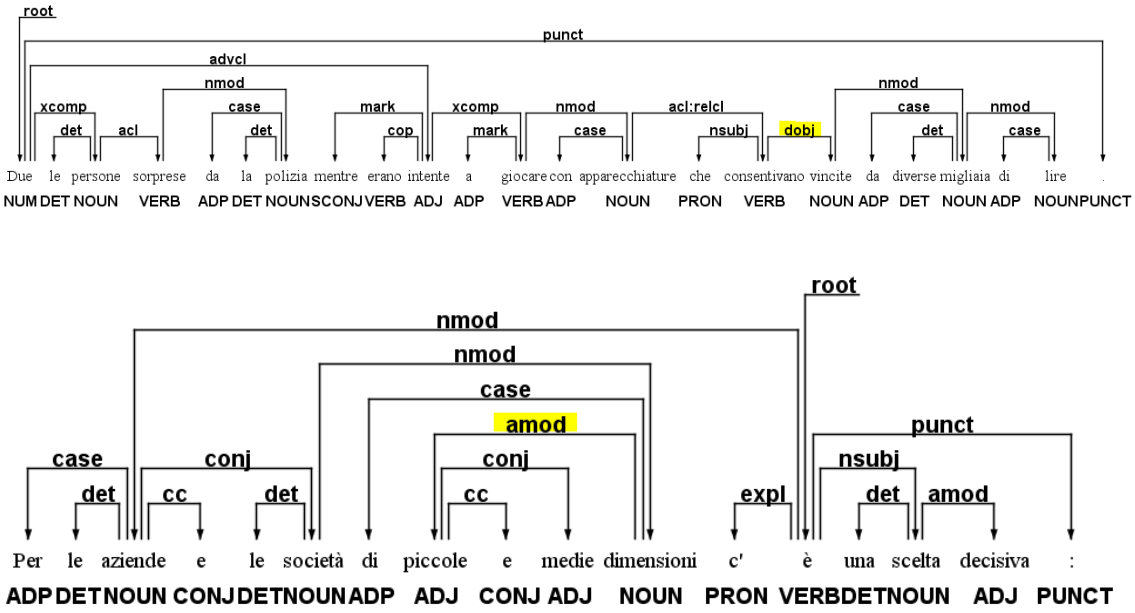


Figura 4.1: Alberi sintattici di frasi estratte dalla treebank italiana. Gli archi che ottengono un basso punteggio sono segnalati in giallo.

La seconda fra le due frasi d'esempio, così come altre all'interno del corpus, può essere interpretata come un fenomeno di marcatezza: le costruzioni marcate, fra le altre cose, possono essere definite come quelle costruzioni che nel linguaggio hanno meno probabilità di realizzarsi rispetto ad altre semanticamente equivalenti. Nel capitolo 2.4 sono state affrontate le diverse posizioni della teoria linguistica riguardo il fenomeno di marcatezza, con particolare attenzione alla definizione che interpreta le costruzioni marcate come deviazione rispetto alla "norma linguistica" caratterizzate da una bassa frequenza d'uso. Si sottolinea nuovamente che in ogni caso la letteratura in materia riconosce una forte correlazione fra marcatezza e complessità. Questo significa che per mezzo di strumenti come LISCA dovrebbe essere possibile intercettare i fenomeni di marcatezza: se da un lato costruzioni non marcate saranno caratterizzate da un basso livello di complessità (dunque un maggior livello di

plausibilità di annotazione), dall'altro costruzioni caratterizzate da gradi crescenti di marcatezza saranno associate a una maggiore complessità (equivalente a punteggi di minore plausibilità). L'intuizione quindi è che LISCA può essere usato per ricostruire il passaggio graduale da costruzioni non marcate a costruzioni caratterizzate da gradi crescenti di marcatezza [Tusa et al., 2016].

Quando ci si trova di fronte ad una frase non marcata, la sua maggiore o minore complessità può essere dovuta a diversi fattori che agiscono su livelli distinti del linguaggio (e.g. piano lessicale, sintattico, etc.) [Fiorentino, 2009] e che rendono la frase più difficile da elaborare anche sul piano cognitivo [Hawkins, 1994]. Dal momento che il punteggio di un arco calcolato da LISCA si basa su parametri strutturali linguisticamente motivati di cui fanno parte anche quelle informazioni che la letteratura individua come cause di complessità linguistica (più ne dettaglio nel capitolo 4.2.1), si può intendere il valore ottenuto come un indice di complessità della costruzione[3].

Si passerà ora alla descrizione dell'algoritmo di LISCA e quali i parametri attraverso cui ottiene i propri punteggi. Si tenga presente che LISCA opera in due fasi: prima colleziona statistiche sulla base di una serie di *features* linguistiche estratte da un ampio corpus di frasi analizzate; in una seconda fase calcola il punteggio per ciascun link di dipendenza sulla base delle statistiche che ha ricavato nel passaggio precedente.

4.2.1 L'algoritmo di LISCA

L'input di LISCA è rappresentato dal risultato dell'analisi sintattica di un parser a dipendenze automatico. Il punteggio di LISCA viene assegnato a ciascun arco di dipendenza del corpus analizzato, definito come una tripla *(d, h, t)* dove d è il dipendente, h è la sua testa sintattica e t è il tipo di relazione di dipendenza che connette d ad h. Prendendo in considerazione sia proprietà strutturali che linguistiche dell'arco, viene calcolato un punteggio qualitativo poi utilizzato per ordinare tutti gli archi di dipendenza del corpus in ordine di affidabilità.

Per facilitare l'interpretazione dell'ordinamento prodotto da LISCA, in questo lavoro si è deciso di suddividere tale ordinamento in dieci fasce di uguali dimensioni, le quali rappresentano i diversi gradi di plausibilità assegnati da LISCA: nella fascia

[3]Si tenga conto di questa nozione perché verrà utilizzata più volte nel corso dell'analisi dei risultati della ricerca, descritti nel capitolo 5.

1 si troveranno gli archi che hanno ottenuto i punteggi più alti, mentre in fascia 10 quelli meno plausibili; le fasce da 2 a 9 sono le fasce intermedie.

Il software è stato progettato per essere indipendente sia rispetto al linguaggio (nella sua variante de-lessicalizzata) che agli strumenti utilizzati nelle fasi preliminari. Questo significa che non sono imposti vincoli sul parser che deve essere usato, né tanto meno sullo standard di annotazione o sulla lingua che si desidera analizzare.

Per gli scopi di questa tesi infatti, che si pone l'obiettivo di osservare i fenomeni linguistici in una prospettiva multilingua, LISCA è stato utilizzato nella sua variante delessicalizzata: questo ha permesso di fare astrazione da variazioni di natura lessicale. Sono invece state prese in considerazione altre caratteristiche e proprietà relative all'arco che riguardano sia le caratteristiche associate dalla letteratura linguistica alla nozione di complessità sintattica, sia che tengono conto dell'albero a dipendenze che include l'arco stesso. Per questa ragione si può dire che LISCA fa riferimento sia a tratti locali che a tratti globali, raffigurati nell'immagine 4.2.

I primi, i tratti locali, si riferiscono alla singola relazione rappresentata dall'arco considerato e fanno riferimento per esempio alla distanza lineare in tokens fra testa e dipendente, alle categorie grammaticali ad essi assegnate in fase di POS tagging, alla loro forza associativa che le unisce o al rapporto fra POS della testa t e tipo di dipendenza. I tratti globali invece, più complessi, sono volti a localizzare l'arco all'interno della struttura sintattica della frase. Trattandosi di tratti più complessi, ma anche più informativi, verranno affrontati a breve nel dettaglio. Si tratta infatti della distanza di d rispetto alla radice dell'albero, alla foglia più vicina o a quella più lontana, oppure ancora il numero di nodi "fratelli" e "figli" di d ricorrenti rispettivamente alla sua destra o sinistra nell'ordine lineare della frase.

I tratti selezionati per il calcolo del punteggio di plausibilità si possono considerare "linguisticamente motivati" in quanto basati sulla struttura dell'albero a dipendenze, ma anche focalizzati su strutture linguistiche che riflettono la complessità delle frasi a livello sintattico e a livello di parsing [Dell'Orletta et al., 2013].

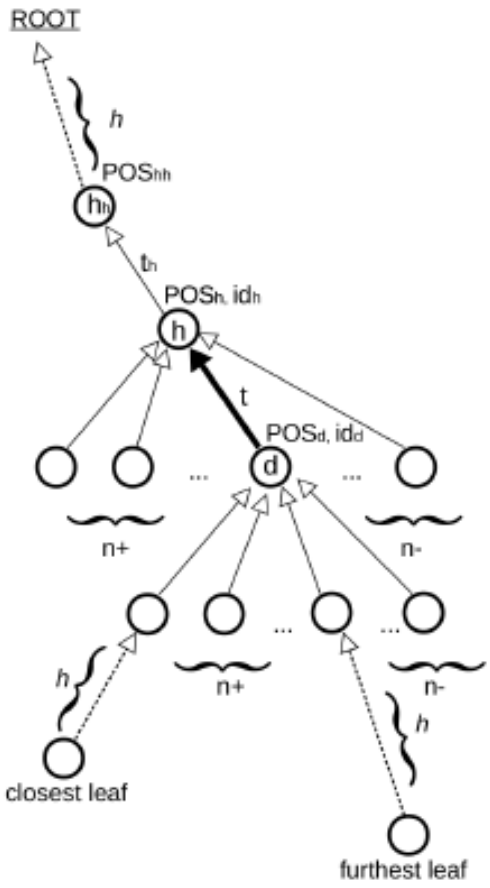


Figura 4.2: Caratteristiche utilizzate LISCA per il calcolo della plausibilità dell'arco *(d; h; t)*.

Plausibilità dell'arco di dipendenza Per plausibilità di un arco di dipendenza si intende uno dei tratti locali calcolato considerando le categorie grammaticali di d, h, e della testa padre di h.

Questo tratto è stato utilizzato per la prima volta in Dell'Orletta et al. [2011] per calcolare l'efficacia dell'analisi di un parser, si è rivelato un indice efficace anche per individuare la correttezza di singoli archi di dipendenza.

Localizzazione di un arco di dipendenza all'interno della struttura globale dell'albero Si passi ora ad osservare i parametri globali presi in considerazione da LISCA.

Localizzare un dato collegamento (o link) di dipendenza all'interno della struttura globale della frase è un'operazione piuttosto complessa poiché tiene conto contempo-

raneamente di molti fattori che co-occorrono nella struttura della frase. In particolare si osservano tre proprietà, relative al dipendente *d*, utili per determinare la posizione di *t*:

- Distanza di *d* dal nodo radice;
- Distanza di *d* dal nodo foglia più vicino;
- Distanza di *d* dal nodo foglia più lontano.

Per ciascuna di queste tre misure viene selezionato il percorso più breve, calcolato sulla base dei nodi che separano i due elementi nella struttura ad albero. Attraverso queste misure, è possibile ricostruire il posizionamento dell'arco all'interno della struttura a dipendenze e dedurre da esse anche la posizione di *t*, in particolare il suo livello di "incassamento" nell'albero.

A questo tratto se ne affiancano altri due che si focalizzano su sotto-alberi a dipendenze locali. Lo scopo di questi tratti è di localizzare *d* tenendo in considerazione l'ordine lineare superficiale delle parole.

Il primo tratto fa riferimento al sotto-albero che dipende da *d.* Di esso si osservano tutti gli immediati dipendenti e, sulla base della posizione che assumono nell'ordine lineare della frase rispetto a *d*, vengono suddivisi in pre- e post-dipendenti. Il secondo prende invece in considerazione i nodi fratelli di *d,* ricostruiti a partire dal sotto-albero governato dalla testa *h* di *d*. A seconda che precedano o seguano *d* nell'ordine lineare delle parole all'interno della frase, anche i nodi fratelli vengono divisi fra pre-dipendenti e post-dipendenti.

Lunghezza e direzione di un arco di dipendenza La lunghezza della dipendenza (abbreviata come DL) rappresenta la distanza lineare tra la testa sintattica *h* e il dipendente *d* in termini di parole intercorrenti. La direzione della dipendenza (d'ora in avanti DD) permette invece a distingue le relazioni in cui la testa precede il dipendente da quelle in cui viceversa lo segue.

Per ogni relazione di dipendenza tra le parole *Wi* e *Wj*, se *Wi* è la testa e *Wj* è il dipendente, allora la lunghezza della dipendenza può essere calcolata come la differenza *i* e *j*. Calcolandola in questo modo, la DL assoluta fra parole adiacenti

collegate da un link di dipendenza risulta pari a 1. Quando i è più grande di j, la DL equivale ad un numero positivo, stando a indicare che la testa si trova dopo il dipendente. Quando invece i è più piccola di j la DL corrisponde a un numero negativo, e nella frase la testa precede il suo dipendente.

La lunghezza di una dipendenza, lontana dall'essere un semplice indice quantitativo, è direttamente proporzionale alla complessità della relazione: più è lunga la dipendenza, più complessa la relazione. "L'ordine relativo delle parole e la loro distanza influenza fortemente la probabilità che una parola ne modifichi un'altra" [Collins, 1996]. In altre parole, più due parole sono vicine, più e probabile che fra esse esista un qualche tipo di relazione.

Il valore di DD rappresenta un misura quantitativa utile nell'analisi di specifiche strutture sintattiche all'interno della collezione di testi: per esempio la posizione della testa rispetto al dipendente in caso di modificatori aggettivali è un indice che permette di operare una distinzione tipologica fra lingue [Liu, 2010]. Più in generale comunque la DD permette di calcolare la proporzione fra costruzioni con testa in posizione iniziale o finale.

Calcolo della qualità dell'arco sintattico Il punteggio qualitativo di un arco di dipendenza (*Quality Score*, QS) viene calcolato come una forza associativa che permette ad un dipendente d di realizzarsi in una certa struttura, tenendo conto della sua testa h e del nodo nonno di d, ovvero 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 et al. [2013], calcola la probabilità dell'arco attraverso il prodotto dei diversi pesi di ciascun tratto. Tali pesi derivano dalle statistiche indotte a partire da un'ampia quantità di dati automaticamente analizzati da un parser a dipendenze che servono a costruire il modello di LISCA. Il modello viene poi applicato ad un nuovo corpus dei cui archi vene calcolata la probabilità.

4.3 Dati e fasi preliminari

La ricerca descritta in questo elaborato è il risultato di un'analisi multilingua sulle treebank UD di italiano, inglese e spagnolo. L'ipotesi di ricerca iniziale prevedeva

che il risultato dello studio avrebbe messo in luce le differenze più significative nella produzione di costruzioni marcate nelle tre lingue, basandosi appunto sulla nozione di marcatezza intesa contemporaneamente come complessità e rarità nei testi. Nei fatti tuttavia la ricerca ha prodotto risultati ben più ricchi e ha catturato molti più fenomeni di quanti fossero attesi, come viene ampiamente descritto nel capitolo 5.

Prima di poter ottenere i punteggi di plausibilità di LISCA è necessario che sui corpora vengano svolte alcune fasi di analisi preliminari. L'intero processo che ha portato dai dati grezzi ai risultati di LISCA può essere sintetizzato in tre fasi:

1. Annotazione sintattica automatica di un corpus di grandi dimensioni per ciascuna delle tre lingue;
2. Estrazione delle statistiche linguistiche e creazione dei modelli grazie all'applicazione di LISCA a ciascuno dei tre corpora annotati;
3. Calcolo dei punteggi di qualità di LISCA per ciascun arco delle treebank gold UD.

La scelta di calcolare i punteggi di LISCA sulle treebank UD è dovuta dal fatto che l'obiettivo della ricerca è quello di ottenere un ordinamento di archi assumibili come corretti. Sebbene anche nelle risorse gold sono contenuti un certo numero di errori, questi dovrebbero essere non particolarmente numerosi; per questa ragione si può concludere che i punteggi di LISCA rappresentano la probabilità che una costruzione si realizzi in una lingua fra tutte quelle grammaticali.

Mentre ai risultati della terza fase è interamente dedicato il capitolo 5, nelle prossime sezioni si descriveranno le prime due fasi e i dati utilizzati per l'addestramento di LISCA.

4.3.1 Annotazione dei corpora

Per il parsing delle risorse linguistiche è stato scelto di realizzare un'annotazione sintattica a dipendenze basata sullo standard delle Universal Dependencies. Volendo applicare LISCA direttamente sulle treebank gold UD, è stato indispensabile servirsi

di una risorsa linguistica sufficientemente grande e rappresentativa per creare i modelli statistici dell'algoritmo. A tale scopo sono stati utilizzati tre corpora provenienti dall'enciclopedia online Wikipedia[4].

Wikipedia è una risorsa multilingua disponibile gratuitamente online costruita grazie al contributo volontario di numerosi collaboratori. La natura collaborativa e lo scopo didascalico che ha mantenuto sin dalla sua nascita, rendono Wikipedia una pratica risorsa linguistica per l'NLP e un buon punto di riferimento per il linguaggio standard. L'ambiente collaborativo infatti appiattisce le differenze linguistiche dettate dagli stili individuali, mentre il genere testuale dell'enciclopedia dovrebbe garantire che tutte le frasi contenute in essa siano grammaticali. Negli ultimi anni si è assistito ad un forte aumento di ricerche svolte su corpora estratti dai social network, in particolare Twitter. Le difficoltà legate a questi generi di testi, che usano un registro colloquiale e un gran numero di costruzioni agrammaticali, dovrebbero essere superate grazie ai testi di Wikipedia, che presentano un registro piuttosto formale e neutrale [Zesch et al., 2007].

I tre corpora monolingue estratti da Wikipedia contengono circa 40 milioni di tokens ciascuno, e sono stati annotati sintatticamente secondo lo schema UD. Le grandi dimensioni dei corpora di addestramento fanno sì che gli errori di annotazione del parser in esse contenuti non abbiamo rilevanza statistica e non compromettano i risultati di LISCA. Tuttavia bisogna comunque tener conto della presenza di tali errori e del dominio di appartenenza dei testi, che sono la ragione per cui LISCA, fra le altre cose, assegna punteggi di plausibilità molto bassi alle costruzioni che sono deviazione rispetto al dominio del corpus di addestramento.

Sebbene ciascuna delle tre lingue presenti delle particolarità, sono disponibili online delle *pipeline* che permettono di utilizzare un parser unico per un ampio numero di lingue. Come è stato già accennato nel capitolo 3.2, il parsing multilingua è un task che attira fortemente l'attenzione della comunità scientifica, tanto che la task condivisa della conferenza CoNLL 2017 riguarda proprio questo tema[5], grazie anche al contributo che il progetto Universal Dependencies ha dato fornendo gli strumenti necessari per lo sviluppo della ricerca in tal senso.

È importante, quando si vuole svolgere una ricerca multilingua, che l'annotazione venga fatta per mezzo di un unico *framework* al fine di evitare inconsistenze di

[4]https://www.wikipedia.org/ (consultato il 07/04/2017).
[5]http://universaldependencies.org/conll17/ (consultato il 07/04/2017).

annotazione. Per raggiungere un tale obiettivo, sul sito della conferenza CoNLL viene suggerito l'utilizzo di due strumenti, entrambi basati sullo standard UD, che si propongono come soluzione al problema. Si tratta di SyntaxNet[6] e UDPipe[7]. Entrambi gli strumenti tuttavia hanno rivelato delle difficoltà legate al loro effettivo utilizzo.

La prima scelta è ricaduta su SyntaxNet [Petrov, 2016] dal momento che i risultati delle sue performance sono stati valutati su 36 lingue diverse e sono migliori rispetto a quelli ottenuti da altri strumenti simili, come dichiarato dagli stessi autori [Andor et al., 2016]. SyntaxNet è un'implementazione open-source di un modello di analisi linguistica integrato in TensorFlow. Le reti neurali sono recentemente diventate una tecnica molto popolare nei task di NLP in virtù dei brillanti risultati che permettono di ottenere nei compiti di *pattern recognition*. SyntaxNet sfrutta proprio questa tecnica. L'utilizzo di questo strumento è vincolato all'uso di modelli di addestramento diversi per ciascuna lingua che permettano di estrarne le particolarità da riapplicare in fase di parsing. Trattandosi comunque di uno strumento pensato per risolvere i problemi legati al parsing multilingua, i modelli usati per addestrare SyntaxNet possono rappresentare qualsiasi lingua. La creazione di un modello tuttavia è un'operazione complessa, ragion per cui ne vengono messi a disposizione 36 pre-addestrati, ciascuno per una lingua diversa, che si basano sullo standard di annotazione UD, per la precisione UD versione 1.3.

L'utilizzo effettivo di SyntaxNet sui corpora estratti da Wikipedia ha però messo in luce i limiti dello strumento. Sebbene l'analisi del corpus inglese risultasse piuttosto accurata, sono emerse delle problematiche legate all'analisi della lingua italiana e spagnola. Una catena di analisi linguistica automatica che prenda in input testo puro deve prevedere di risolvere anche le fasi precedenti la vera e propria annotazione che, se non realizzate correttamente, possono compromettere il risultato finale dell'analisi. La suddivisione del testo in frasi prima e la tokenizzazione poi sono operazioni fondamentali per permettere una corretta analisi linguistica. Il tokenizzatore implementato in SyntaxNet invece non realizza la separazione dei pronomi clitici dal verbo cui sono legati (i.e. "vederci", la cui tokenizzazione corrisponde a "vedere" e separatamente "ci") e di preposizioni articolate (i.e. la tokenizzazione di "dello" si realizza nei due token distinti "del" e "lo"). Dal momento che il sistema che dovrebbe operare la separazione del testo in parole sintattiche non gestisce cor-

[6] https://www.tensorflow.org/versions/r0.11/tutorials/syntaxnet/ (consultato il 07/04/2017).
[7] https://ufal.mff.cuni.cz/udpipe (consultato il 07/04/2017).

rettamente questi casi, ad essi viene assegnata un'errata parte del discorso; questo fatto compromette il risultato dell'analisi sintattica. Il tokenizzatore implementato in SyntaxNet infatti nasce con lo scopo di ricevere in input solo frasi della lingua inglese, che tuttavia non presenta elementi linguistici come i pronomi clitici e le preposizioni articolate, di cui invece sono ricchi italiano e spagnolo. Gli alti punteggi di accuratezza ottenuti da SyntaxNet sono infatti relativi al solo parsing linguistico (classi grammaticali e dipendenze sintattiche) e non fanno riferimento ai risultati che si ottengono utilizzando SyntaxNet per l'intera catena di analisi. Dovendo in ogni caso ricercare uno strumento alternativo per la separazione del testo in parole sintattiche, si è preferito non utilizzare SyntaxNet per nessuna delle tre lingue.

Si è scelto quindi di testare la seconda delle *pipeline* suggerite dalla conferenza CoNLL, ovvero UDPipe [Straka et al., 2016]. Si tratta anche in questo caso di un sistema basato sulle reti neurali addestrato sul modello Universal Dependencies, in questo caso versione 1.2. I risultati di questa pipeline sembrano ottenere valori di accuratezza addirittura superiori rispetto a SyntaxNet, ma comunque comparabili[8]. Anche per questo strumento tuttavia sono emerse delle problematiche, nuovamente per quanto riguarda la tokenizzazione della lingua italiana. Mentre in spagnolo, che presenta sia l'uso dei pronomi clitici che delle preposizioni articolate, non si sono riscontrate difficoltà, lo stesso non si può dire per l'italiano: per tali casi viene sistematicamente realizza una lemmatizzazione errata e una conseguentemente sbagliata assegnazione della parte del discorso. Anche UDPipe, come SyntaxNet, è uno strumento che fornisce un *framework* unico per l'analisi linguistica indipendente dalla lingua; lo strumento infatti può, come SyntaxNet, essere addestrato con qualsiasi modello. A differenza del precedente, tuttavia, gli autori di UDPipe hanno sviluppato internamente al sistema anche gli strumenti per tokenizzazione e separazione del testo in frasi. Anch'essi sono addestrati sul modello delle Universal Dependencies, cosa che li rende più performanti rispetto a quelli che sono stati riadattati per il precedente strumento. Ad ogni modo sia SyntaxNet che UDPipe si sono rivelati inadeguati per l'analisi della lingua italiana, per cui la suddivisione del testo in frasi, parole sintattiche e relativo POS tagging di tale lingua sono stati svolti grazie ad uno strumento alternativo, ovvero ILC-POS-Tagger [Dell'Orletta, 2009], che ha un'accuratezza del 96,34% nell'identificazione simultanea della categoria grammaticale e dei

[8]de Lhoneux et al. [2016] confrontano i valori di LAS ottenuti da UDPipe, MaltParser e SyntaxNet. Per l'inglese, ad esempio, i valori ottenuti dai tre sistemi rispettivamente sono 81.3%, 79.9% e 80.38%. Sebbene in questo caso UDPipe ottenga risultati migliori di SyntaxNet, nell'articolo viene sottolineato che i risultati di quest'ultimo non sono direttamente comparabili perché utilizzano un diverso POS tagger e analizzatore morfologico.

tratti morfologici associati. Il testo così analizzato è stato verticalizzato e passato ad UDPipe per svolgere l'analisi sintattica.

Il risultato di questa prima fase sono tre corpora annotati sintatticamente secondo lo standard Universal Dependencies (versione 1.2) restituiti nel formato conll-u. I corpora sono rappresentativi ciascuno del linguaggio standard della propria lingua, ovvero italiano, inglese e spagnolo.

4.3.2 LISCA e treebank della ricerca

La seconda fase prevede l'estrazione delle statistiche linguistiche a partire dai tre corpora ottenuti dalla fase precedente. Tali corpora sono stati usati per l'addestramento di LISCA e la creazione dei modelli, ragion per cui hanno dimensioni fra loro comparabili: circa 40 milioni di tokens ciascuno.

Una volta creati i modelli (sulla base dei parametri descritti nella sezione 4.2), LISCA è stato applicato direttamente sulle treebank gold di UD (anch'esse versione 1.2), ciascuna lingua sul modello corrispondente. Questo processo ha assegnato un valore di plausibilità ad ogni arco della risorsa. La plausibilità di un arco in questo contesto deve essere intesa come un valore che indica quanto è probabile che una certa costruzione si realizzi nel linguaggio rispetto a tutte le costruzioni grammaticali della lingua semanticamente equivalenti ad essa.

L'output di LISCA è stato infine disposto secondo un ordine di plausibilità decrescente e gli archi sono stati raggruppati in 10 fasce che contengono circa il 10% degli archi totali del corpus ciascuna. Di seguito la tabella riporta il dettaglio di ciascuna lingua.

	N di archi totali	N di archi non zero	N di archi per fascia
Inglese	231827	228578 (98,6%)	22857
Italiano	261424	260621 (99,6%)	26062
Spagnolo	426520	426313 (99,9%)	42631

Tabella 4.1: Riassunto delle caratteristiche strutturali delle tre treebank

A differenza di italiano e inglese, che hanno dimensioni simili, la treebank UD dello spagnolo ha un numero di archi maggiore rispetto alle altre due. Per risolvere queste disomogeneità, i risultati di LISCA vengono sempre normalizzati sotto forma di percentuali rispetto al totale. In questo modo si possono confrontare anche i risultati di corpora di dimensioni diverse.

Grazie ai risultati di LISCA è possibile ottenere diverse informazioni sul comportamento delle tipologie di archi della treebank. Per esempio è possibile osservare la distribuzione per fascia degli archi che hanno una determinata relazione di dipendenza; oppure, più sofisticata, la distribuzione degli archi che hanno una determinata relazione di dipendenza e che hanno una certa categoria grammaticale come testa della relazione.

Queste sono solo alcune delle informazioni che sono state ricavate grazie ai risultati di LISCA e di cui si parlerà nel dettaglio nel prossimo capitolo.

Capitolo 5

Analisi dei risultati

In questo capitolo verranno presentati i risultati delle analisi condotte sulla treebank italiana, spagnola e inglese.

Come prima cosa si offrirà una panoramica sulla distribuzione delle classi grammaticali e delle relazioni di dipendenza attraverso le fasce di LISCA. Successivamente si passerà ad analizzare un'informazione fortemente vincolata alla tipologia della lingua, ovvero l'ordine lineare dei costituenti. In ultima battuta si presenteranno nel dettaglio alcune relazioni di dipendenza.

5.1 Distribuzione delle Part-Of-Speech

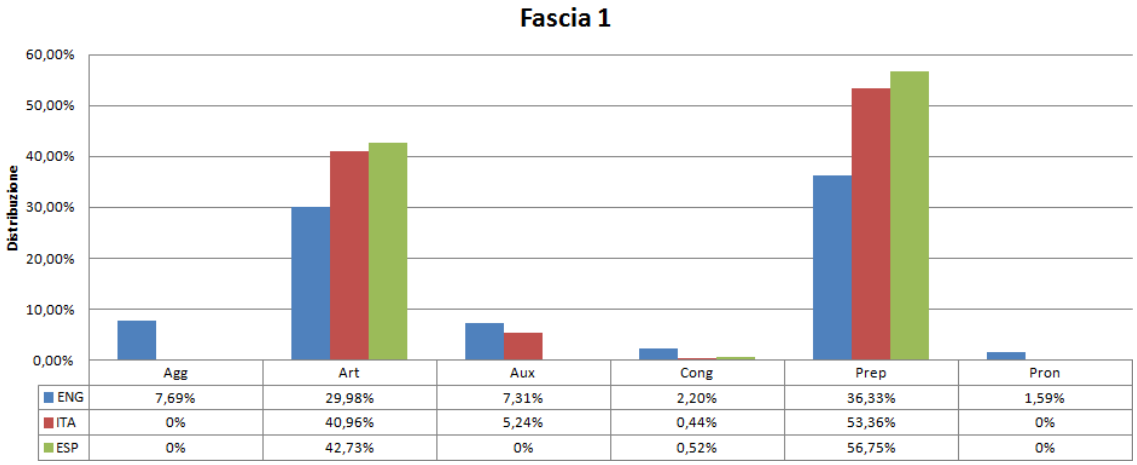
I primi dati riguardano le classi grammaticali (anche *part-of-speech*, quindi POS) e il modo in cui si distribuiscono all'interno delle diverse fasce nelle quali sono organizzati i punteggi di LISCA.

Uno dei principi fondamentali del progetto Universal Dependencies è quello di fornire uno schema di annotazione e delle linee guida per il suo utilizzo che siano universalmente valide per qualsiasi lingua naturale. I *tag* che rappresentano le classi grammaticali in UD sono stati pensati in modo tale che qualsiasi lingua potesse usufruirne, anche servendosi di classi sotto-specificate per le costruzioni caratteristiche di un numero ristretto di lingue. Per ampliare ulteriormente l'applicabilità, non viene imposto l'obbligo di adoperare tutte le etichette messe a disposizione dallo

schema. In uno studio multilingua si deve tenere conto di questi fattori per far sì che le treebank siano fra loro confrontabili. Sono state fatte quindi delle scelte di esclusione o accorpamento di determinate classi.

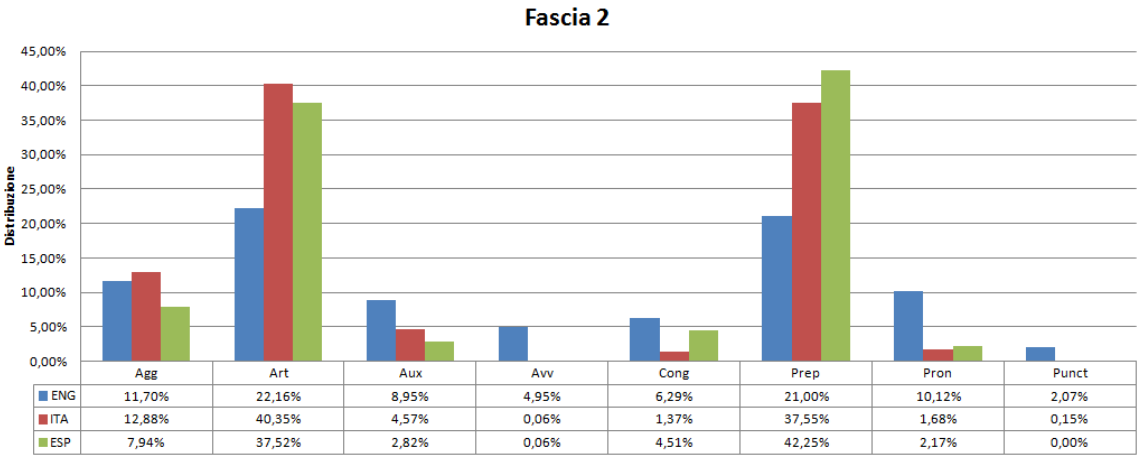
Le congiunzioni, con l'eliminazione della distinzione fra subordinanti e coordinanti, ricadono fra le POS che hanno subito il processo di accorpamento, mentre sono state del tutto escluse le particelle marcate come PART (*particle*). Quest'ultima è un'etichetta usata quasi esclusivamente nella treebank inglese per i casi di genitivo sassone (*"-'s"*) e per la preposizione *"to"* quando precede un verbo alla forma infinito. Il *tag* è comunque presente anche in italiano e spagnolo, seppur raro. In italiano per esempio sono PART i casi di nomi propri che contengono un genitivo sassone (e.g. *McDonand's*). In spagnolo invece si usa questa etichetta per prefissi come *"ex-"*, *"vice-"*, *"trans-"*, etc.

Fra le POS comuni alle tre treebank ce ne sono alcune che non risultano linguisticamente informative e che solitamente non vengono indagate. Si tratta in questo caso di nomi propri (PROPN), simboli (SYM), interiezioni (INTJ), parole straniere (X) e numeri (NUM). Infine sono state escluse anche tutte quelle categorie che nella fascia non raggiungevano una soglia di occorrenza minima, stabilita all'1%, poiché non dotate di alcuna rilevanza statistica data la loro bassa frequenza.

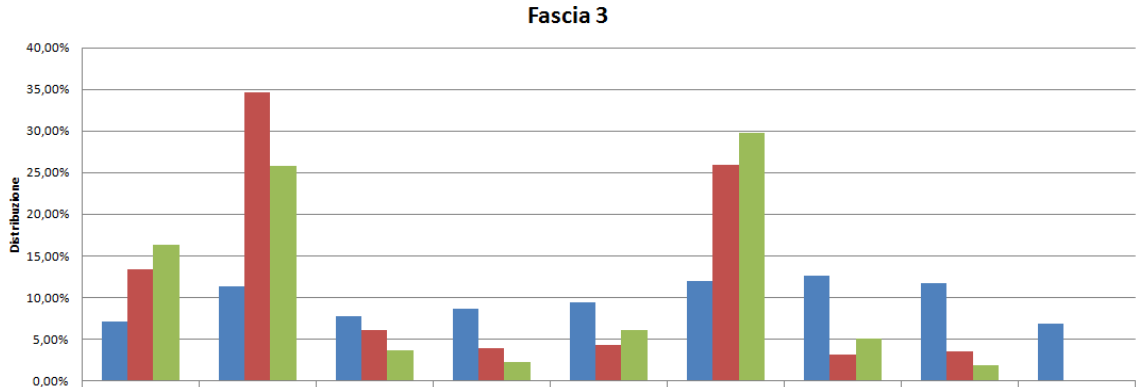


	Agg	Art	Aux	Cong	Prep	Pron
ENG	7,69%	29,98%	7,31%	2,20%	36,33%	1,59%
ITA	0%	40,96%	5,24%	0,44%	53,36%	0%
ESP	0%	42,73%	0%	0,52%	56,75%	0%

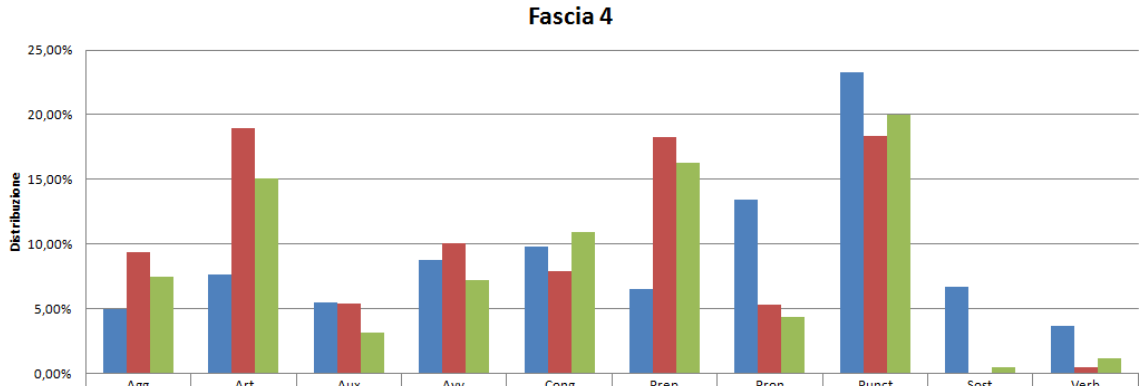
Figura 5.1: Distribuzione delle POS. Fascia 1.



	Agg	Art	Aux	Avv	Cong	Prep	Pron	Punct
ENG	11,70%	22,16%	8,95%	4,95%	6,29%	21,00%	10,12%	2,07%
ITA	12,88%	40,35%	4,57%	0,06%	1,37%	37,55%	1,68%	0,15%
ESP	7,94%	37,52%	2,82%	0,06%	4,51%	42,25%	2,17%	0,00%

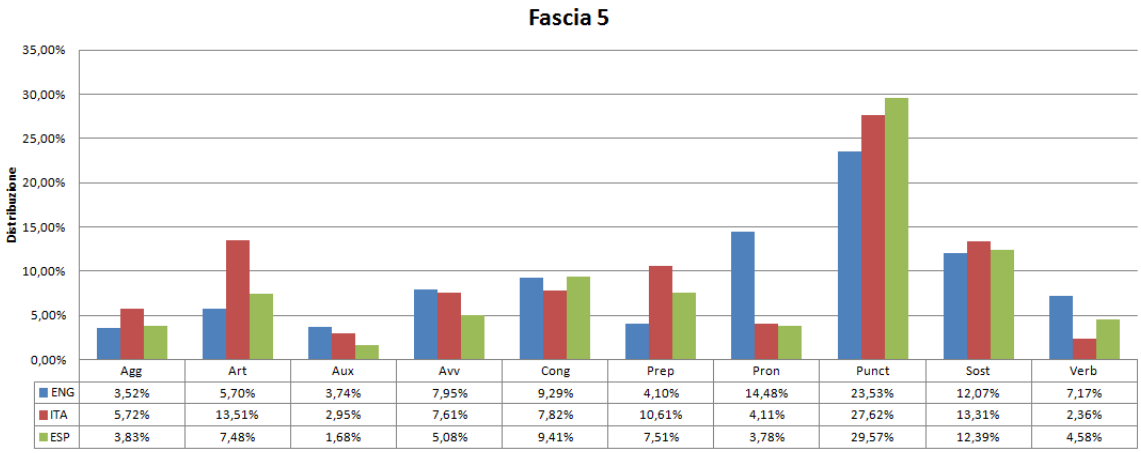


	Agg	Art	Aux	Avv	Cong	Prep	Pron	Punct	Sost
ENG	7,12%	11,30%	7,74%	8,60%	9,44%	12,02%	12,66%	11,75%	6,89%
ITA	13,40%	34,60%	6,03%	3,85%	4,24%	25,92%	3,20%	3,50%	0,00%
ESP	16,38%	25,83%	3,59%	2,28%	6,10%	29,83%	5,11%	1,81%	0,00%

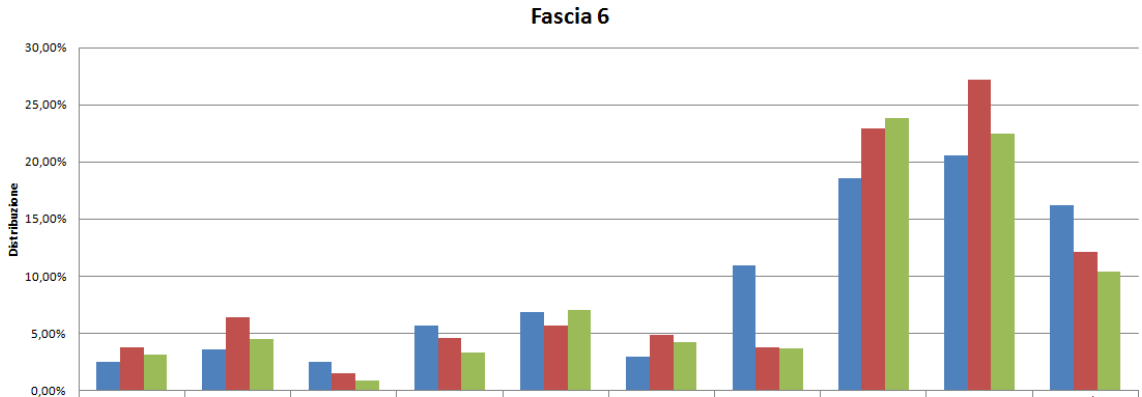


	Agg	Art	Aux	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	4,96%	7,65%	5,53%	8,74%	9,79%	6,49%	13,47%	23,27%	6,69%	3,69%
ITA	9,37%	18,93%	5,36%	10,06%	7,89%	18,31%	5,35%	18,39%	0,00%	0,46%
ESP	7,49%	15,05%	3,17%	7,25%	10,90%	16,28%	4,36%	20,03%	0,49%	1,16%

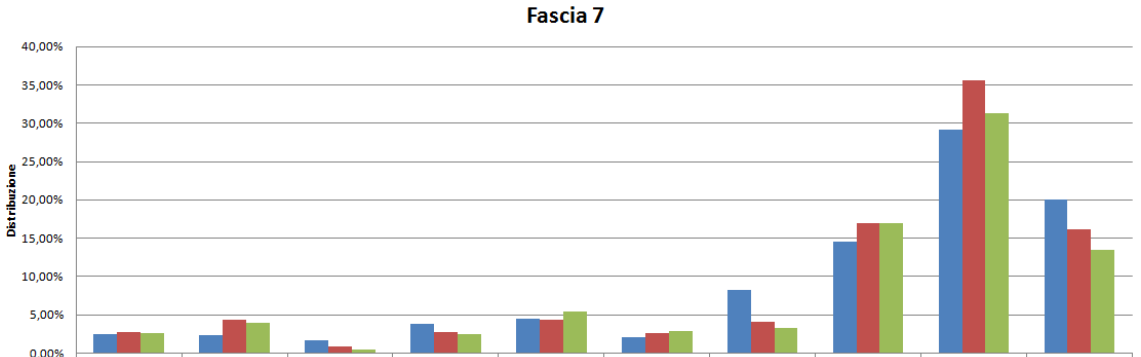
Figura 5.2: Distribuzione delle POS. Fasce 2-4.



	Agg	Art	Aux	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	3,52%	5,70%	3,74%	7,95%	9,29%	4,10%	14,48%	23,53%	12,07%	7,17%
ITA	5,72%	13,51%	2,95%	7,61%	7,82%	10,61%	4,11%	27,62%	13,31%	2,36%
ESP	3,83%	7,48%	1,68%	5,08%	9,41%	7,51%	3,78%	29,57%	12,39%	4,58%

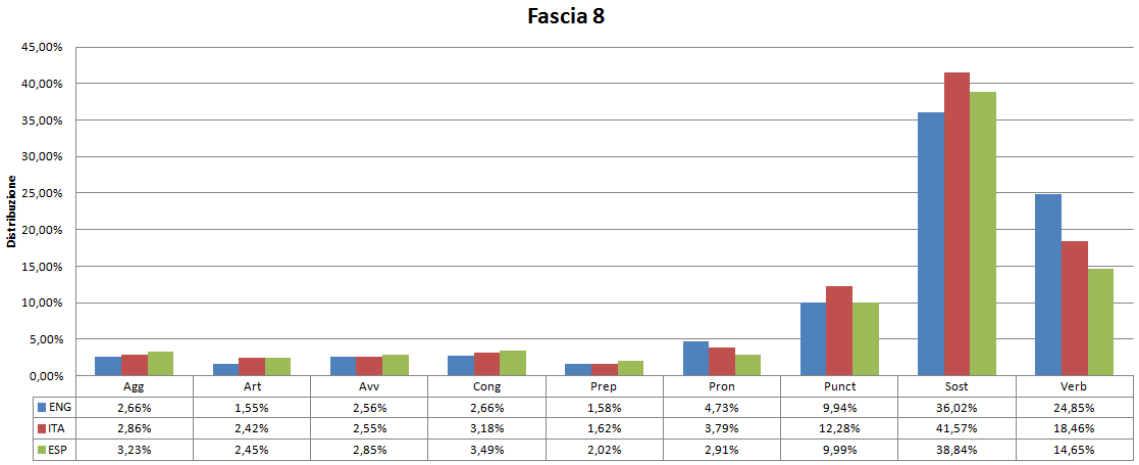


	Agg	Art	Aux	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	2,53%	3,55%	2,52%	5,70%	6,89%	2,95%	10,99%	18,56%	20,58%	16,19%
ITA	3,76%	6,44%	1,45%	4,59%	5,66%	4,88%	3,79%	22,89%	27,23%	12,08%
ESP	3,09%	4,47%	0,83%	3,35%	7,03%	4,20%	3,72%	23,87%	22,45%	10,35%

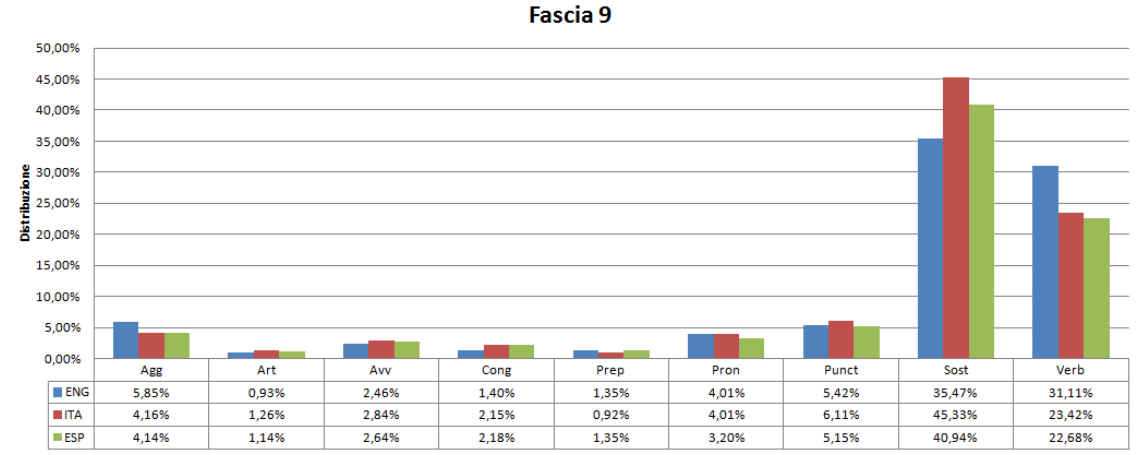


	Agg	Art	Aux	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	2,53%	2,38%	1,64%	3,87%	4,52%	2,07%	8,20%	14,52%	29,13%	20,01%
ITA	2,80%	4,39%	0,89%	2,71%	4,37%	2,59%	4,06%	16,97%	35,57%	16,13%
ESP	2,55%	3,95%	0,51%	2,52%	5,41%	2,91%	3,25%	16,99%	31,34%	13,47%

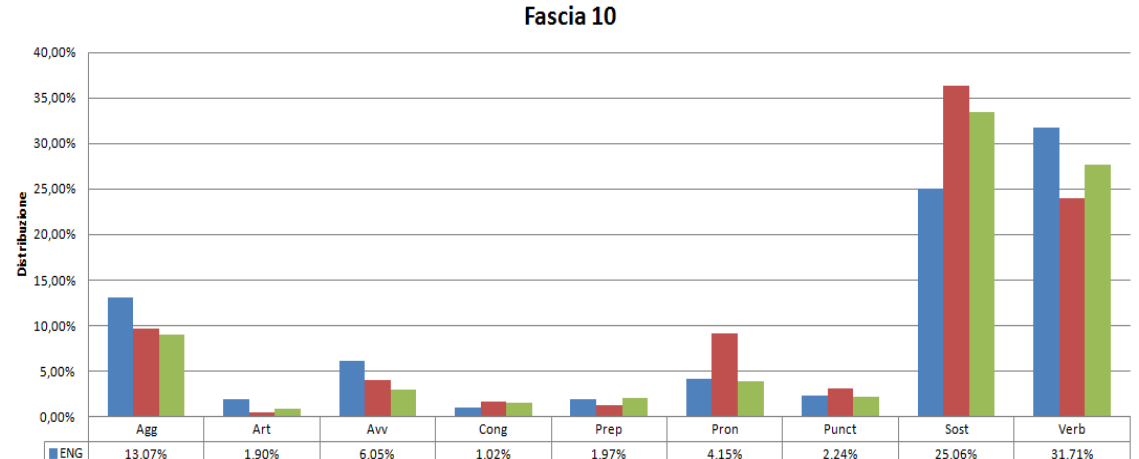
Figura 5.3: Distribuzione delle POS. Fasce 5-7.



	Agg	Art	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	2,66%	1,55%	2,56%	2,66%	1,58%	4,73%	9,94%	36,02%	24,85%
ITA	2,86%	2,42%	2,55%	3,18%	1,62%	3,79%	12,28%	41,57%	18,46%
ESP	3,23%	2,45%	2,85%	3,49%	2,02%	2,91%	9,99%	38,84%	14,65%



	Agg	Art	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	5,85%	0,93%	2,46%	1,40%	1,35%	4,01%	5,42%	35,47%	31,11%
ITA	4,16%	1,26%	2,84%	2,15%	0,92%	4,01%	6,11%	45,33%	23,42%
ESP	4,14%	1,14%	2,64%	2,18%	1,35%	3,20%	5,15%	40,94%	22,68%



	Agg	Art	Avv	Cong	Prep	Pron	Punct	Sost	Verb
ENG	13,07%	1,90%	6,05%	1,02%	1,97%	4,15%	2,24%	25,06%	31,71%
ITA	9,63%	0,48%	3,99%	1,61%	1,19%	9,12%	3,06%	36,35%	23,97%
ESP	9,01%	0,86%	2,90%	1,50%	2,07%	3,88%	2,15%	33,38%	27,61%

Figura 5.4: Distribuzione delle POS. Fasce 8-10.

Osservando i grafici sulla distribuzione delle classi grammaticali attraverso le fasce di LISCA è possibile distinguere quali siano le POS più complesse e quali quelle più semplici sulla base dell'ordinamento prodotto dall'algoritmo[1].

Preposizioni ed articoli si collocano nelle fasce più alte (1-3), mentre sostantivi e verbi in quelle più basse (7-10), e questo accade in tutte le lingue. È possibile spiegare questo dato pensando al modo in cui queste classi entrano in relazione con gli altri elementi della frase e al modo in cui si costituiscono: le preposizioni e gli articoli sono parti del discorso invariabili che rispettano delle regole di comportamento piuttosto rigide e regolari indipendentemente dagli elementi con cui si combinano, invece sostantivi e verbi risentono maggiormente del contesto in cui sono inseriti. Queste caratteristiche rendono alcune classi grammaticali più rigide e regolari di altre e, di conseguenza, più semplici per LISCA.

Oltre alla distribuzione nelle diverse fasce delle classi grammaticali, su cui si ritornerà in seguito indagando alcune categorie mirate, quello che è importante mettere in luce è il fatto che le tre lingue presentano dei comportamenti generali piuttosto diversi. La lingua inglese per esempio si distingue da italiano e spagnolo per il numero di classi che appaiono in ogni singola fascia, in particolar modo in quelle iniziali: gli elementi della lingua inglese che si collocano nella prima fascia si distribuiscono su un numero maggiore di POS rispetto a italiano e spagnolo, per cui invece ne appaiono solo due.

L'inglese è una lingua dotata di una morfologia flessionale più semplice rispetto alle lingue romanze (si veda capitolo 2.6); contrariamente ad esse per esempio non è dotata di morfemi di genere (i.e. maschile e femminile). LISCA risente di questa caratteristica rintracciando dei modelli di comportamento ricorrenti in un numero maggiore di classi grammaticali. Ciò può essere interpretato come un primo indizio del fatto che i dati ricavati grazie a LISCA permettono di distinguere le lingue sulla base della loro tipologia e della famiglia linguistica.

Un'ulteriore prova di questo fatto viene fornita dalle distribuzioni delle POS in italiano e spagnolo, che sono sostanzialmente sovrapponibili. La vicinanza fra queste due lingue fa sì che i fenomeni linguistici delle rispettive treebank siano molto simili, discostandosi da quelli dell'inglese.

[1]La semplicità o complessità di un elemento in questo contesto sono intese in funzione dei punteggi calcolati da LISCA, come discusso nel capitolo 4.2 (pagina 46).

Oltre alle differenze è tuttavia possibile rintracciare anche un certo numero di somiglianze fra le tre lingue. È questo il caso per esempio della distribuzione delle parole piene e delle classi funzionali.

In linguistica si definiscono parole piene tutte quelle parole che sono dotate di un significato semantico proprio, mentre appartengono alle classi funzionali quelle parole che svolgono solamente funzioni grammaticali e non possiedono una semantica indipendente. Sono parole piene i sostantivi, i verbi, gli aggettivi e gli avverbi; mentre sono categorie funzionali le preposizioni, gli articoli, gli ausiliari e le congiunzioni. Questa distinzione coincide con quella che si realizza fra classi grammaticali aperte e classi grammaticali chiuse: la proprietà di essere aperte o chiuse descrive la possibilità che all'interno della classe si inseriscano nuovi elementi. Difatti se da una parte è vero che in una lingua vengono continuamente generati nuovi sostantivi, le preposizioni di una lingua tendono ad essere stabili e non modificarsi nel tempo [Berruto and Cerruti, 2011].

I grafici in figura 5.5 mostrano chiaramente che, per quanto riguarda la distribuzione per fascia delle classi grammaticali aperte e chiuse, le tre lingue hanno pari andamento: le prime fasce sono occupate quasi esclusivamente da elementi appartenenti alle classi grammaticali chiuse, mentre nelle ultime la tendenza si inverte e le classi aperte diventano le più numerose. Le parole piene sono più flessibili, ammettono maggiore mobilità all'interno della frase e sono morfologicamente più ricche, ragioni per cui si collocano nelle ultime fasce di LISCA.

La distinzione della linguistica tradizionale in classi grammaticali aperte e chiuse trova riscontro nelle distribuzioni di LISCA. Il fatto che il principio sia valido per tutte e tre le lingue fa pensare che potrebbe trattarsi di un principio universale: le classi aperte costituiscono un tipo di elemento morfosintattico più complesso rispetto alle classi funzionali. La complessità in questo caso si può intendere sotto diversi aspetti: oltre a quelli già affrontati che riguardano struttura e morfologia, anche dal punto di vista lessicale le classi aperte sono più ricche. Esperimenti di neurolinguistica infatti hanno dimostrato che nell'elaborazione di parole appartenenti alle classi aperte l'area di attivazione neurale è più estesa rispetto a quella delle classi chiuse; coinvolge per esempio anche l'area visiva, assente nel caso dell'elaborazione delle classi chiuse [Mohr et al., 1994].

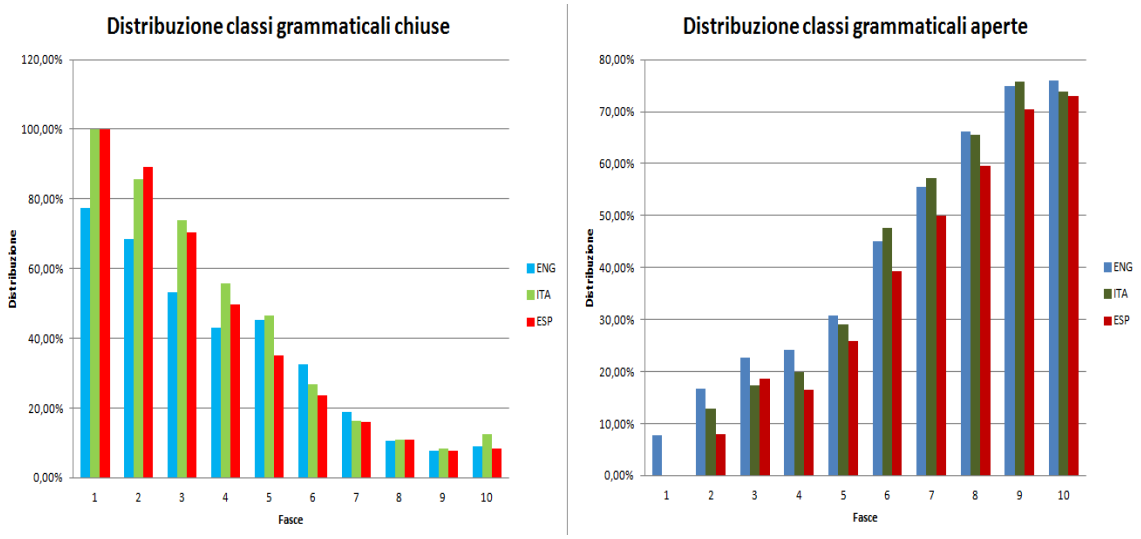


Figura 5.5: Distribuzione (percentuale) delle classi grammaticali chiuse e aperte per fascia.

Per poter delineare un quadro più dettagliato del comportamento delle tre lingue, sono state messe sotto la lente di ingrandimento alcune categorie grammaticali la cui distribuzione può fornire informazioni interessanti.

Articoli e preposizioni per esempio si collocano prevalentemente nelle fasce dalla 1 alla 4, dopodiché il loro andamento si stabilizza su valori molto bassi. Anomalo è invece il comportamento dei pronomi inglesi. Le altre categorie presentano degli andamenti piuttosto simili in tutte e tre le lingue o hanno delle differenze che emergono solo grazie allo studio delle corrispondenti relazioni sintattiche, per cui quindi si rimanda alla sezione successiva (capitolo 5.4).

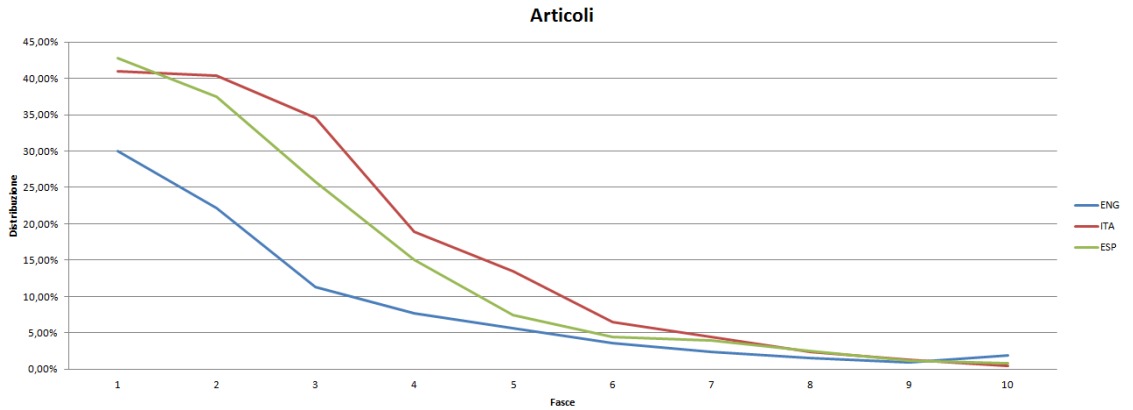


Figura 5.6: Distribuzione (percentuale) degli elementi DET per fascia.

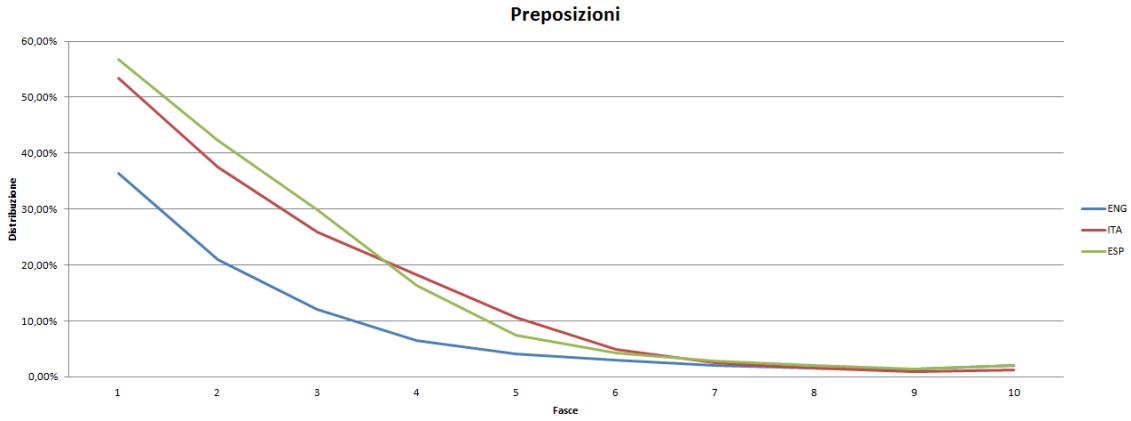


Figura 5.7: Distribuzione (percentuale) degli elementi ADP per fascia.

Articoli e preposizioni I grafici che rappresentano la distribuzione degli articoli e delle preposizioni mostrano come le due classi rispettino la stessa tendenza, indipendentemente dalla treebank in cui appaiono (figure 5.6 e 5.7).

I valori dell'inglese sono tuttavia inferiori rispetto a italiano e spagnolo. Potrebbe quindi sembrare che la lingua inglese faccia minore uso di queste classi funzionali, e di conseguenza all'interno della treebank se ne osservino in quantità più ridotta. Allo stesso tempo però è ben più probabile che questa discrepanza di valori derivi piuttosto da scelte di annotazione specifiche della lingua: se accompagna un verbo all'infinito, la preposizione *"to"* viene classificata come PART nella treebank inglese, mentre non accade lo stesso per nessuna delle preposizioni italiane e spagnole. Si tratta di una annotazione ereditata dalla Penn Treebank e se ne osservano numerosi casi nonostante le linee guida di UD non diano indicazioni in tal senso.

Nel confronto diretto fra le lingue i contenuti delle classi non risultano perfettamente confrontabili e quello che emerge sono le discrepanze nei criteri di annotazione. Si tratta comunque di informazioni utili, non tanto in una prospettiva linguistica quanto piuttosto in un'ottica di revisione o omogeneizzazione dei dati. Questo è in linea con i possibili casi d'uso di LISCA che coinvolgono sia l'analisi linguistica che la revisione delle treebank.

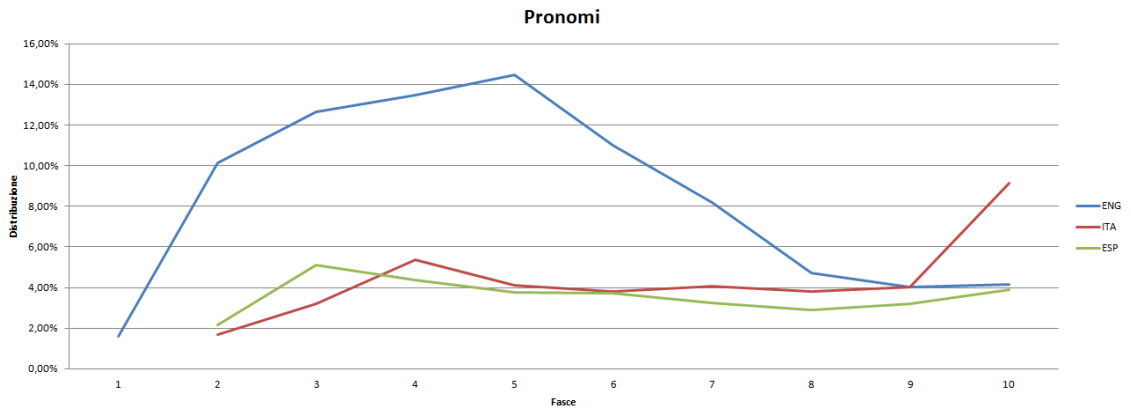


Figura 5.8: Distribuzione (percentuale) degli elementi PRON per fascia.

Pronomi L'andamento dei pronomi è fra tutti il più sorprendente, ma anche facilmente interpretabile se si osserva il contenuto dei raggruppamenti.

Nella grammatica tradizionale inglese è prevista una categoria, chiamata *personal*, di cui fanno parte i pronomi personali, i pronomi possessivi e i determinanti possessivi [Halliday and Hasan, 1976]. Di fatto però i determinanti possessivi svolgono nella frase la funzione di aggettivi: si tratta appunto delle forme "*my*", "*your*", "*our*", "*their*", etc., che per una scelta di annotazione specifica della lingua inglese vengono considerati pronomi (PRON), a differenza delle altre lingue in cui sono determinanti (DET) con l'aggiunta di una *feature* che ne indica il valore di possessività (*Poss=Yes*).

È naturale dunque che gli aggettivi possessivi abbiamo un'altissima influenza sull'andamento dei pronomi nella treebank inglese: innanzitutto si tratta di una classe ad alto tasso di frequenza all'interno della risorsa, ed inoltre la loro struttura sintattica è piuttosto rigida, fattore che li concentra nelle prime fasce.

5.2 Distribuzione delle dipendenze

In questa sezione verranno analizzate le relazioni di dipendenza sintattica mostrandone le distribuzioni per fascia e, successivamente, i tratti rilevanti di alcune di esse.

Come già successo per le parti del discorso, anche in questo caso i grafici rappresentano solo una selezione di dati ottenuti dall'analisi, ovvero solo quelli dotati di

rilevanza. Per esempio sono state escluse le relazioni di dipendenza che ottenevano una frequenza per fascia inferiore ad 1%, trattandosi di casi isolati e statisticamente poco rilevanti.

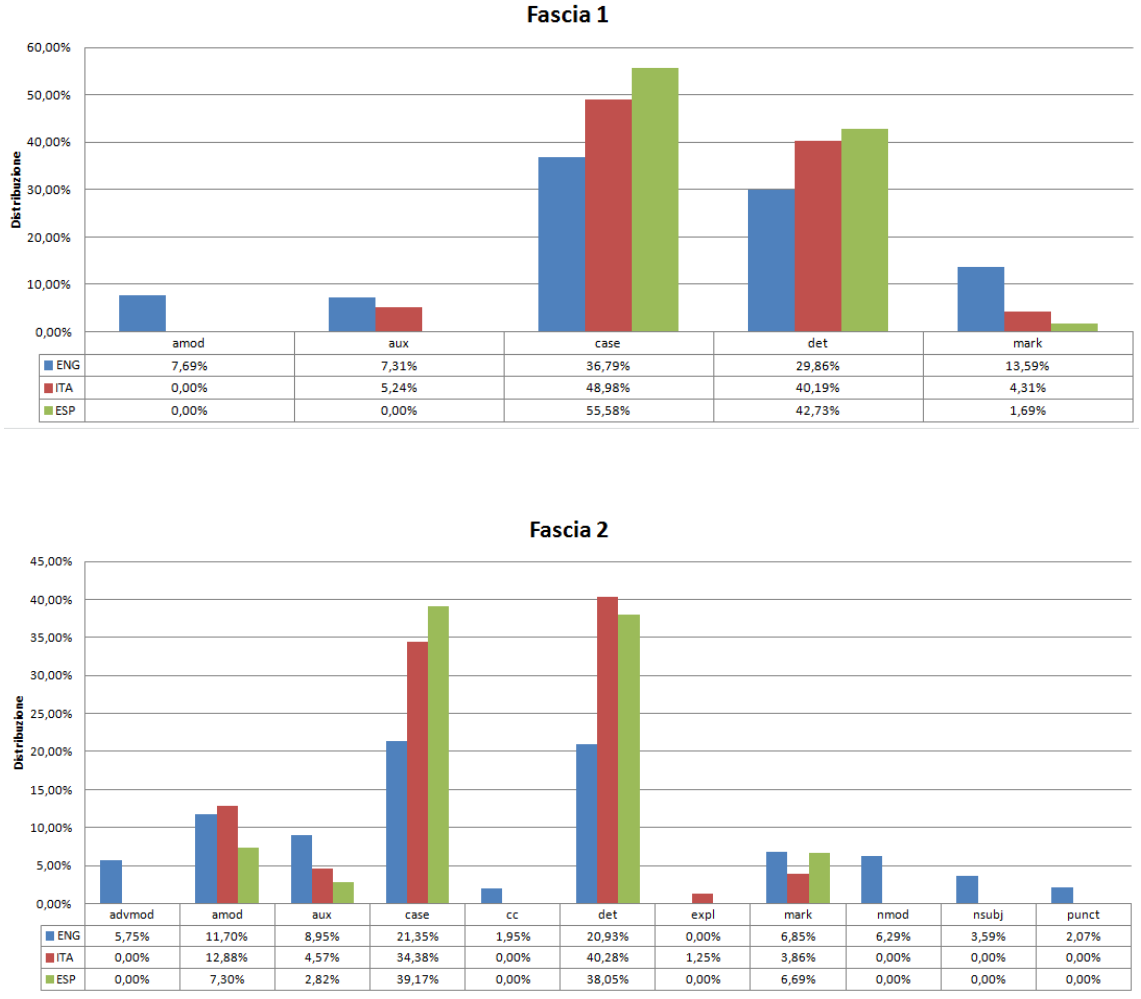
Altre relazioni ancora sono state accorpate fra loro, come accaduto nel caso delle dipendenze per il soggetto e il soggetto passivo: in UD è possibile distinguere i due casi per mezzo di un tag sotto-specificato. Tuttavia il soggetto passivo è un'etichetta che viene utilizzata solo in italiano; per questa ragione, sebbene i casi d'uso delle due relazioni siano diversi, i loro valori sono stati sommati. Stessa cosa si dica per gli ausiliari e gli ausiliari passivi: si è preferito accorpare i valori delle due relazioni e dare una visione d'insieme del fenomeno degli ausiliari. Una scelta simile è stata adottata per le reazioni di dipendenza *nmod* e *nmod:poss*: la prima è l'etichetta utilizzata per i modificatori nominali, mentre la seconda ne rappresenta un caso particolare in cui un modificatore nominale risulta accompagnato dal genitivo sassone *"-'s"*; questa costruzione però è rilevante solo nella risorsa inglese. Le Universal Dependencies prevedono anche una marcatura specifica nel caso di avverbi di negazione ma, dal momento che questi ottengono sempre valori molto bassi, i modificatori avverbiali e gli avverbi negativi sono qui riportati in soluzione unica.

Inoltre sono state totalmente omesse quelle relazioni che vengono adottate per casi specifici di una delle tre lingue e per cui mancherebbe quindi il metro di confronto. Si tratta per esempio della relazione che rappresenta i nomi composti (in UD *compound*). La lingua inglese fa largo uso della giustapposizione fra parole come processo generativo per crearne di nuove, mentre questo non succede altrettanto frequentemente in italiano e spagnolo. Nelle ultime due l'uso di questa relazione è molto ristretto a casi particolari: nella treebank italiana per esempio appare solitamente in ambito giornalistico, e non supera mai i valori di soglia. La presenza di questi archi nella treebank è comunque un dato interessante, dal momento che mette in luce un atteggiamento specifico della lingua inglese che non si riscontra nelle lingue romanze.

Anche la relazione che rappresenta l'oggetto indiretto supera i valori minimi di soglia solo nella treebank spagnola. Questa volta non si tratta di una differenza linguistica bensì di un diverso criterio di annotazione: vengono infatti marcati come *iobj* i pronomi clitici, che invece in italiano sono annotati come oggetti diretti (*dobj*), particelle espletive (*expl*) o anche oggetti indiretti in base alla funzione che svolgono nella frase. Anche in questo caso, come già successo per i pronomi, LISCA ha messo in luce una disomogeneità nei criteri di annotazione delle risorse. Non potendo scio-

gliere gli oggetti indiretti nelle due relazioni corrispondenti per renderli confrontabili con l'italiano, la relazione *iobj* è stata esclusa totalmente dai grafici.

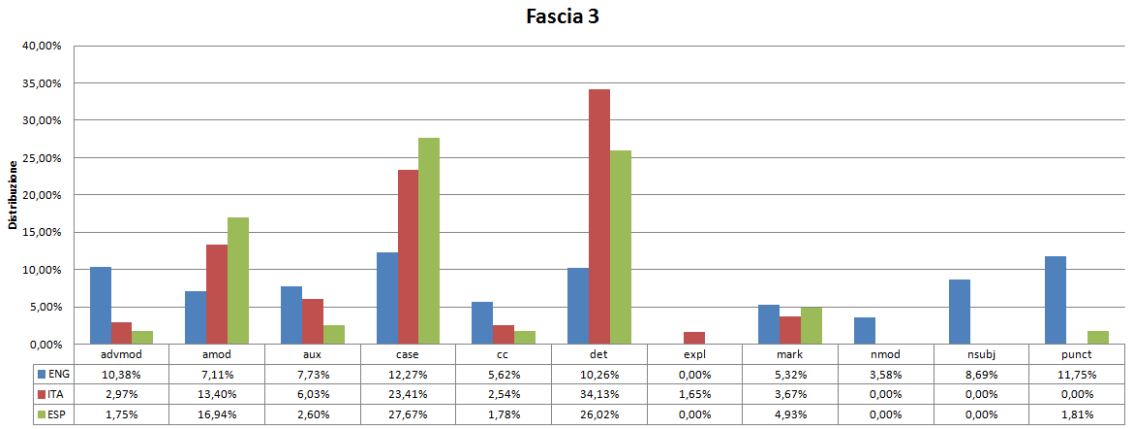
Ci sono infine alcune relazioni marginali dal punto di vista linguistico o che hanno delle frequenze per fascia piuttosto ridotte. Si tratta delle relazioni: *parataxis* (giustapposizione fra elementi), *nummod* (modificatori nominali), *name* (nomi propri) e *mwe* (espressioni multi-parola). Su queste ultime in particolare la comunità di UD si è spesa nel tentativo di delineare dei criteri non ambigui per l'assegnazione di queste relazioni nella nuova versione del tagset, dal momento che nelle release fatte fino ad ora se ne è osservato un uso improprio in tutte le treebank.



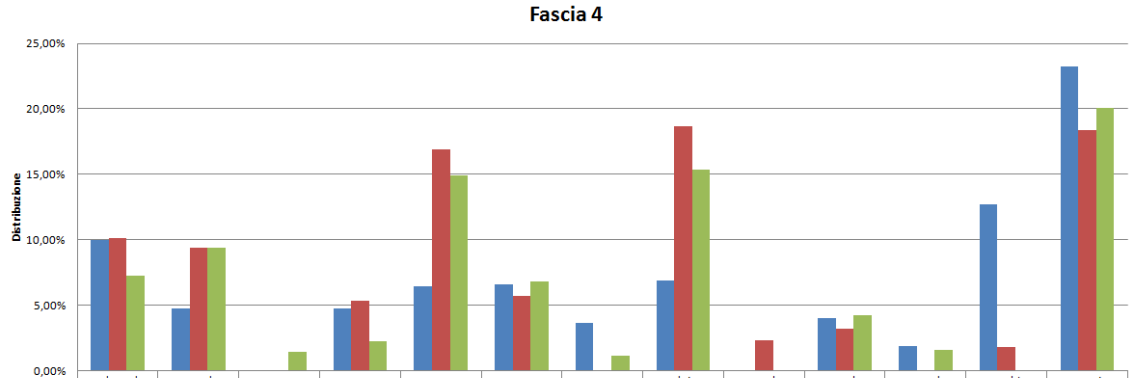
Fascia 1	amod	aux	case	det	mark
ENG	7,69%	7,31%	36,79%	29,86%	13,59%
ITA	0,00%	5,24%	48,98%	40,19%	4,31%
ESP	0,00%	0,00%	55,58%	42,73%	1,69%

Fascia 2	advmod	amod	aux	case	cc	det	expl	mark	nmod	nsubj	punct
ENG	5,75%	11,70%	8,95%	21,35%	1,95%	20,93%	0,00%	6,85%	6,29%	3,59%	2,07%
ITA	0,00%	12,88%	4,57%	34,38%	0,00%	40,28%	1,25%	3,86%	0,00%	0,00%	0,00%
ESP	0,00%	7,30%	2,82%	39,17%	0,00%	38,05%	0,00%	6,69%	0,00%	0,00%	0,00%

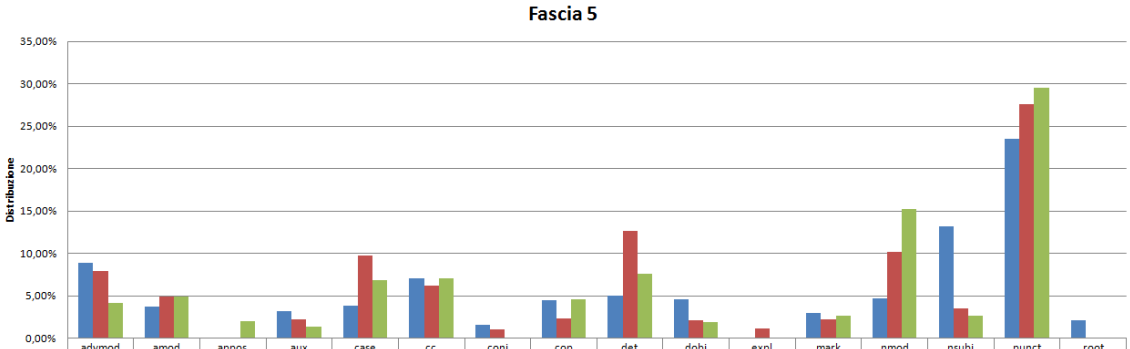
Figura 5.9: Distribuzione relazioni di dipendenza. Fasce 1-2.



	advmod	amod	aux	case	cc	det	expl	mark	nmod	nsubj	punct
ENG	10,38%	7,11%	7,73%	12,27%	5,62%	10,26%	0,00%	5,32%	3,58%	8,69%	11,75%
ITA	2,97%	13,40%	6,03%	23,41%	2,54%	34,13%	1,65%	3,67%	0,00%	0,00%	0,00%
ESP	1,75%	16,94%	2,60%	27,67%	1,78%	26,02%	0,00%	4,93%	0,00%	0,00%	1,81%



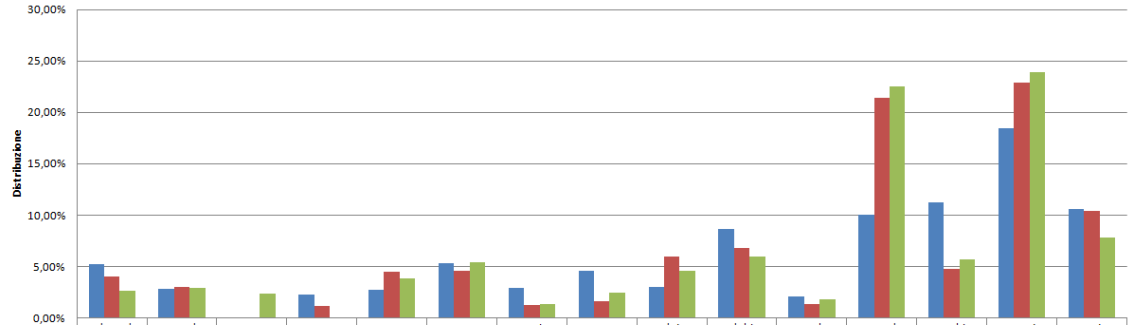
	advmod	amod	appos	aux	case	cc	cop	det	expl	mark	nmod	nsubj	punct
ENG	10,00%	4,72%	0,00%	4,72%	6,48%	6,61%	3,62%	6,85%	0,00%	4,02%	1,86%	12,73%	23,26%
ITA	10,16%	9,37%	0,00%	5,36%	16,94%	5,69%	0,00%	18,66%	2,31%	3,20%	0,00%	1,78%	18,39%
ESP	7,23%	9,42%	1,44%	2,27%	14,93%	6,84%	1,16%	15,36%	0,00%	4,24%	1,55%	0,00%	20,04%



	advmod	amod	appos	aux	case	cc	conj	cop	det	dobj	expl	mark	nmod	nsubj	punct	root
ENG	8,89%	3,75%	0,00%	3,20%	3,78%	7,00%	1,53%	4,49%	4,96%	4,52%	0,00%	2,97%	4,71%	13,18%	23,48%	2,11%
ITA	7,95%	4,92%	0,00%	2,18%	9,71%	6,19%	1,02%	2,36%	12,61%	2,08%	1,12%	2,16%	10,14%	3,47%	27,62%	0,00%
ESP	4,13%	4,89%	1,96%	1,32%	6,81%	7,00%	0,00%	4,53%	7,58%	1,83%	0,00%	2,62%	15,19%	2,61%	29,57%	0,00%

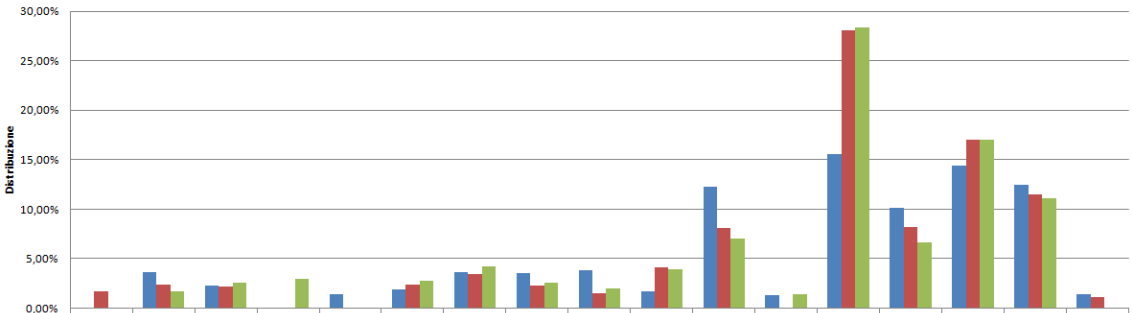
Figura 5.10: Distribuzione relazioni di dipendenza. Fasce 3-5.

Fascia 6



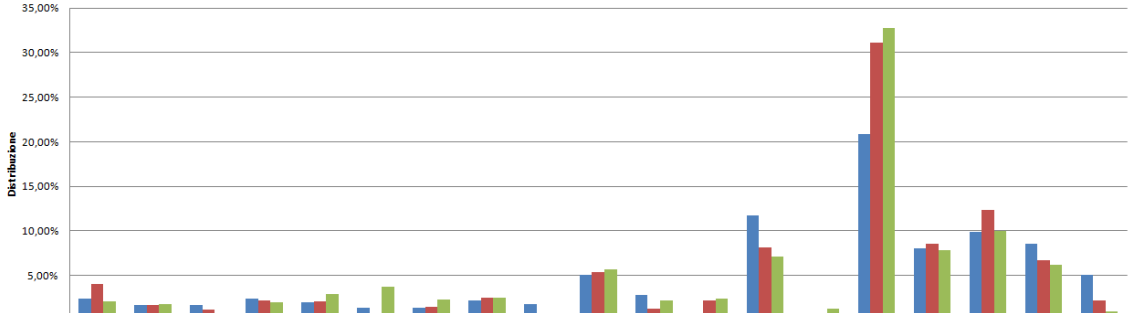
	advmod	amod	appos	aux	case	cc	conj	cop	det	dobj	mark	nmod	nsubj	punct	root
ENG	5,22%	2,80%	0,00%	2,30%	2,70%	5,31%	2,88%	4,54%	2,98%	8,62%	2,06%	10,03%	11,23%	18,49%	10,56%
ITA	3,98%	3,00%	0,00%	1,12%	4,50%	4,54%	1,26%	1,64%	6,00%	6,82%	1,34%	21,44%	4,75%	22,89%	10,42%
ESP	2,66%	2,90%	2,35%	0,00%	3,88%	5,39%	1,35%	2,49%	4,54%	5,93%	1,76%	22,49%	5,69%	23,91%	7,83%

Fascia 7



	acl	advmod	amod	appos	aux	case	cc	conj	cop	det	dobj	mark	nmod	nsubj	punct	root	xcomp
ENG	0,00%	3,60%	2,30%	0,00%	1,44%	1,89%	3,62%	3,57%	3,82%	1,73%	12,22%	1,36%	15,56%	10,13%	14,40%	12,49%	1,43%
ITA	1,70%	2,38%	2,17%	0,00%	0,00%	2,39%	3,48%	2,23%	1,47%	4,13%	8,11%	0,00%	28,03%	8,17%	16,97%	11,46%	1,12%
ESP	0,00%	1,74%	2,58%	2,95%	0,00%	2,79%	4,26%	2,62%	1,99%	3,95%	7,06%	1,43%	28,34%	6,65%	17,00%	11,10%	0,00%

Fascia 8



	acl	acl:relcl	advcl	advmod	amod	appos	case	cc	ccomp	conj	cop	det	dobj	mark	nmod	nsubj	punct	root	xcomp
ENG	2,36%	1,74%	1,68%	2,38%	2,00%	1,35%	1,40%	2,16%	1,82%	5,10%	2,85%	0,00%	11,73%	0,00%	20,89%	8,02%	9,88%	8,61%	5,09%
ITA	4,09%	1,69%	1,15%	2,17%	2,11%	0,00%	1,45%	2,56%	0,00%	5,41%	1,24%	2,25%	8,13%	0,00%	31,15%	8,53%	12,32%	6,71%	2,17%
ESP	2,09%	1,77%	0,00%	1,96%	2,89%	3,73%	2,30%	2,53%	0,00%	5,73%	2,17%	2,46%	7,12%	1,28%	32,76%	7,89%	10,00%	6,20%	1,00%

Figura 5.11: Distribuzione delle relazioni di dipendenza. Fasce 6-8.

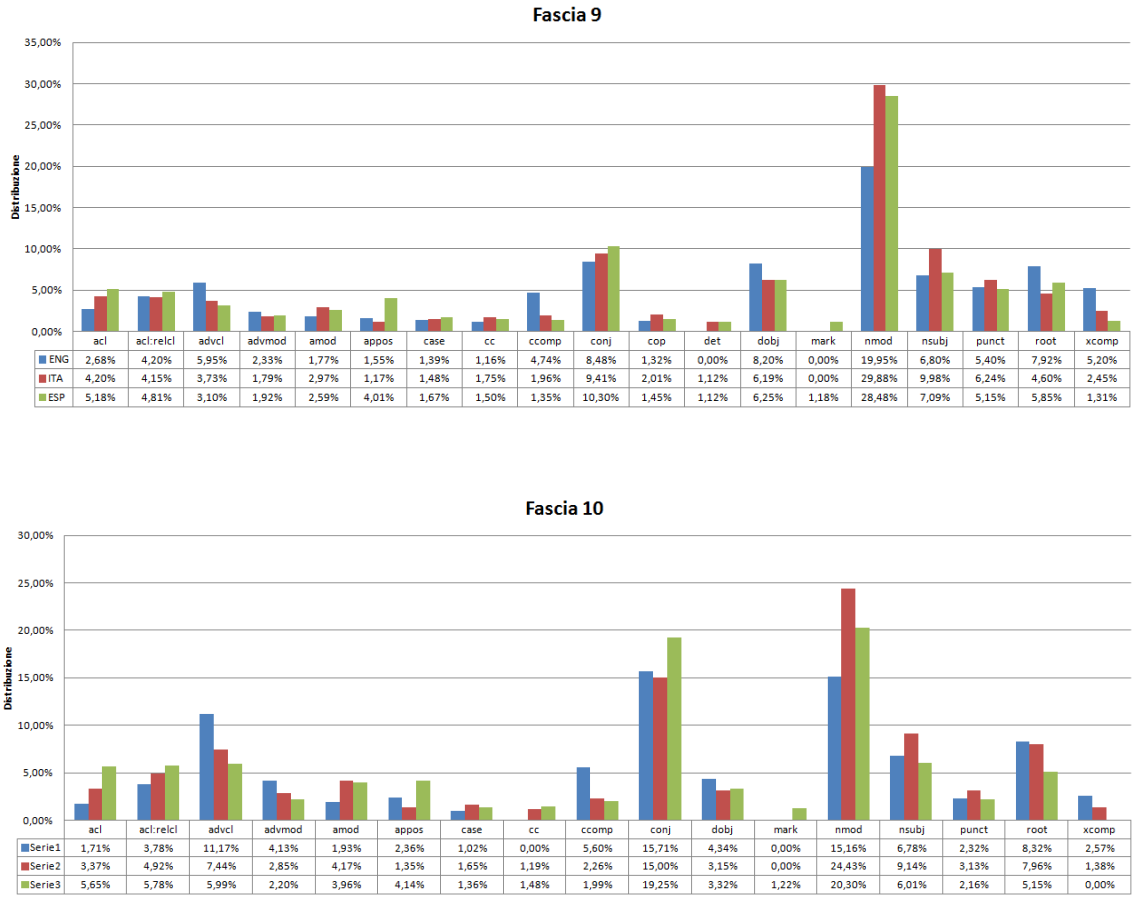


Figura 5.12: Distribuzione relazioni di dipendenza. Fasce 9-10.

Le classi grammaticali degli elementi impongono dei vincoli sul modo in cui i due possono interagire sintatticamente fra loro; ad esempio la relazione *aux* (*auxiliary*) non può che realizzarsi fra un predicato e il suo verbo ausiliare. Per questo motivo non deve sorprendere il fatto che la distribuzione delle relazioni di dipendenza rispecchi in parte quella delle POS, in particolare nelle prime fasce. Gli elementi dell'inglese anche in questo caso provengono da un numero maggiore di relazioni sintattiche, cosa che rende le prime fasce di questa lingua più variegate rispetto a quelle delle altre due, che anche in questi grafici tendono ad avere comportamenti molto simili. In effetti per tutte e tre le lingue l'andamento delle relazioni di dipendenza rispecchia quello registrato per le categorie grammaticali: 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. Si ricorda che la plausibilità di un arco in questo contesto è intesa, come detto nel capitolo 4,

come la probabilità che la relazione che esso rappresenta si realizzi nel linguaggio.

Coerentemente con quanto rilevato per le POS, le relazioni più semplici ordinate da LISCA riguardano le classi grammaticali chiuse, ovvero articoli e preposizioni, rappresentate da *case* e *det*. Fra le preposizioni, quelle che introducono nomi sono considerate più prevedibili rispetto a quelle che accompagnano i verbi. La presenza precoce di archi *aux* conferma la precedente intuizione che si tratti di verbi che si caratterizzano per un comportamento rigido e regolare.

Le relazioni che LISCA segnala come più complesse sono quelle relative alle proposizioni subordinate, che in UD si realizzano come *acl*, *acl:relcl*, *advcl*, *xcom* e *ccomp*. In queste relazioni il dipendente è un predicato che svolge la funzione di radice per un sottoalbero della frase. Il loro grado di complessità è sostanzialmente pari in tutte e tre le lingue e deriva dal fatto che si tratta di frasi annidate all'interno dell'albero sintattico.

Prima di passare ad una analisi più dettagliata di alcune relazioni più significative, è bene spendere alcune parole sul comportamento delle congiunzioni (*conj* e *cc*) e della punteggiatura (*punct*). La punteggiatura era già stata discussa nel capitolo 3.1 poiché si tratta di uno di quei casi in cui la teoria linguistica non può venire in aiuto nella definizione dei criteri di annotazione, e il caso della punteggiatura è analogo. Diversamente da come si potrebbe immaginare infatti queste due relazioni si collocano prevalentemente nelle ultime fasce. L'uso della punteggiatura è di fatto molto soggettivo e infatti oggetto di studio centrale per molti linguisti [Garavelli, 2014]. Nelle risorse linguistiche annotate si aggiunge a questo anche la mancanza di criteri di annotazione omogenei e coerenti, non solo fra lingue diverse ma anche all'interno della stessa lingua. La relazione *punct* infatti non viene mai presa in considerazione nelle fasi di valutazione di sistemi di parsing a dipendenze [Nivre, 2016]. Per quanto riguarda le congiunzioni, si riporta un esempio di costruzioni equivalenti a cui è stata assegnata diversa annotazione. Nel primo esempio, la congiunzione e l'elemento congiunto si fanno dipendere entrambi dal verbo che nell'ordine lineare appare per primo, nel secondo caso avviene il contrario.

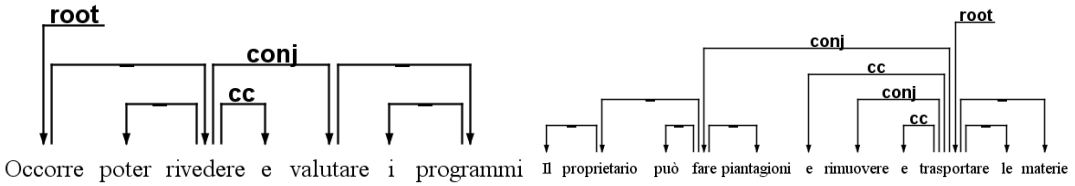


Figura 5.13: Due possibili rappresentazioni della congiunzione coordinante.

Sebbene esistano delle linee guida fornite da UD per il trattamento delle relazioni *cc* e *conj*, di fatto nelle treebank si possono rintracciare entrambe le modalità di trattamento. Nella nuova versione di UD, rilasciata a marzo 2017, per la punteggiatura è stato fatto un tentativo di uniformazione dei trattamenti attraverso una revisione delle linee guida per il suo utilizzo, ovvero imponendo che la testa della relazione sia costituita da una parola piena. La stessa cosa non è avvenuta per le congiunzioni per cui, come si è visto, si possono osservare diverse annotazioni basate su diversi principi che, seppur formalmente corrette, creano disomogeneità nella risorsa.

5.3 Ordinamento delle dipendenze

Un parametro particolarmente rilevante dal punto di vista tipologico per stabilire la prototipicità di una relazione è l'orientamento delle strutture sintattiche, ricavato dall'osservazione della direzione degli archi dell'albero a dipendenze. Tale direzione viene definita in base ai due possibili orientamenti (destra e sinistra) che le relazioni di dipendenza possono assumere nell'ordine lineare della frase: a seconda che la testa di una costruzione sintattica si trovi a destra o a sinistra rispetto al suo dipendente, la direzione dell'arco che li collega sarà orientata verso la testa della relazione.

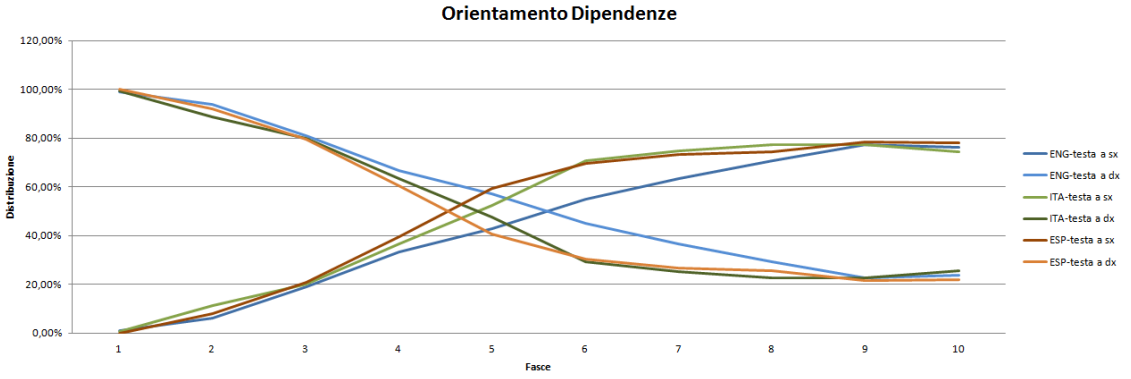


Figura 5.14: Distribuzione (percentuale) degli archi delle treebank in base alla posizione del dipendente rispetto alla testa.

Il grafico mostra come i due tipi di orientamenti (destra e sinistra), nonostante occorrano con una frequenza simile, vengano descritti da andamenti opposti: nelle prime fasce si concentrano le costruzioni con testa a destra (es. "la mamma", dove "mamma" si trova a destra del suo determinante), nelle ultime quelle con testa a sinistra

(es. "mangia la pasta", dove la testa verbale si trova a sinistra). Si noti come l'andamento dei due fenomeni per le tre lingue sia identico: nelle prime fasce si trovano le relazioni con testa a sinistra, che man mano che si procede nelle fasce si esauriscono e vengono sostituite dalle relazioni con testa a destra. Questo dato è da ricondursi 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 per i determinanti) presentano un ordinamento fisso e dunque computativamente più prevedibile.

L'orientamento delle relazioni di dipendenza è un tratto fortemente influenzato dal tipo sintattico cui appartiene una lingua. Per esempio nel capitolo 2.2.2 si è detto come l'ordine dei costituenti fondamentali (soggetto, verbo e complemento oggetto) influiscano le altre relazioni di modo che una lingua risulti prevalentemente pre- o post-determinante.

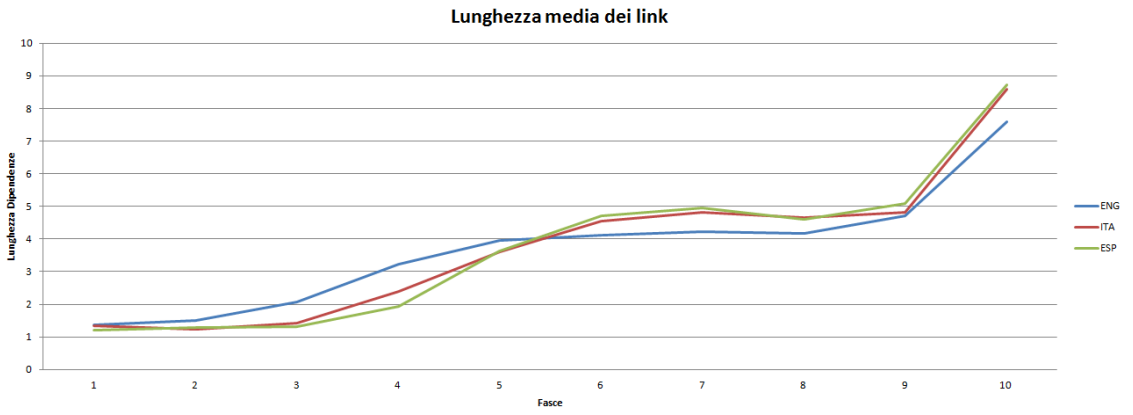


Figura 5.15: Lunghezza media (in parole) fra testa e dipendente.

La distanza in termini lineari di un elemento dalla sua testa sintattica costituisce un altro importante indice di complessità: maggiore la distanza lineare (in parole) fra testa e dipendente, maggiore la complessità di interpretazione dell'informazione fra esse contenuta [Lin, 1996, Gibson, 1998]. Coerentemente con questa nozione di psicolinguistica, LISCA ordina i link più lunghi fra quelli più complessi. Come si può notare dal grafico infatti la lunghezza media degli archi delle tre lingue nelle fasce 1, 2 e 3 non supera il valore 2, mentre subisce un'impennata nella fascia 10. Il fatto che l'andamento sia molto simile per tutte e tre le lingue e non si notino differenze significative dimostra che questa nozione di complessità della frase è universalmente valida a livello inter-linguistico.

L'ordine dei costituenti e la distanza lineare fra essi sono due parametri fortemente correlati: il principio di Dependency Length Minimisation (DLM) [Gildea and Temperley, 2010], stabilisce che se in una lingua sono possibili più ordinamenti per la stessa frase, allora viene preferito quello che minimizza la distanza lineare fra i costituenti. Questo principio è valido universalmente, anche per le lingue romanze che per loro natura ammettono una certa variabilità nell'ordine degli elementi. La maggiore vicinanza fra le parole riduce la quantità di memoria che deve essere impiegata per il processamento della frase e favorisce una comunicazione più immediata; per questa ragione tale principio risulta essere applicato dalle lingue sincronicamente e diacronicamente [Gulordava and Merlo, 2015].

5.4 Le relazioni di dipendenza

Si passerà ora all'analisi di alcune relazioni specifiche. Si tratta di soggetto, complemento oggetto diretto, modificatori aggettivali e avverbiali e infine proposizioni subordinate. Queste relazioni sono state selezionate poiché gli archi che le rappresentano costituiscono secondo Nivre [2016] validi elementi di valutazione dell'accuratezza di un parser.

5.4.1 Soggetto

Il soggetto è una funzione sintattica fondamentale della frase insieme a predicato e oggetto. L'elemento linguistico concreto che ricopre tale funzione ha proprietà diverse nelle varie lingue e può essere individuato da marche morfologiche o sintattiche: il caso nominativo, l'accordo col verbo, la posizione nella frase sono alcune di esse.

Italiano e spagnolo sono lingue a possibile soggetto nullo, mentre nell'inglese questo ruolo è obbligatoriamente espresso[2]. Questo comporta che il numero totale di relazioni *nsubj* presenti nelle tre treebank sia molto diverso: dovendo rendere il soggetto esplicito per ogni predicato, la risorsa inglese ha un numero complessivo di soggetti più elevato rispetto alle altre due. Il maggior numero di esempi, combinato con la regolarità dei soggetti costituiti da pronomi personali, fa sì che gli *nsubj* inglesi ottengano punteggi più alti.

[2]Si veda capitolo 2.6.1 per la differenza fra lingue *pro-drop* e *non pro-drop*.

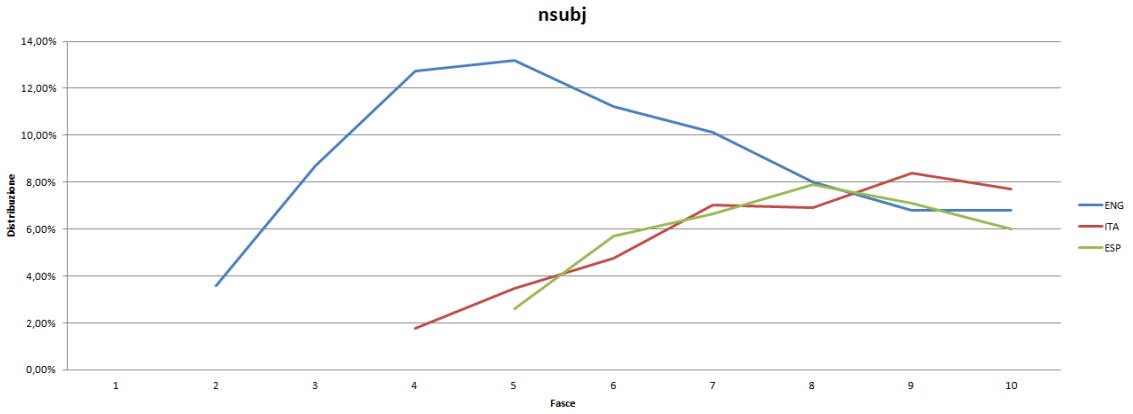


Figura 5.16: Distribuzione (percentuale) di relazioni *nsubj* per fascia.

I due grafici sottostanti rappresentano le possibili posizioni del soggetto rispetto alla propria testa: la dislocazione a destra o sinistra permette di stabilire il focus della frase e fornisce un'indicazione sulla marcatezza o meno della costruzione.

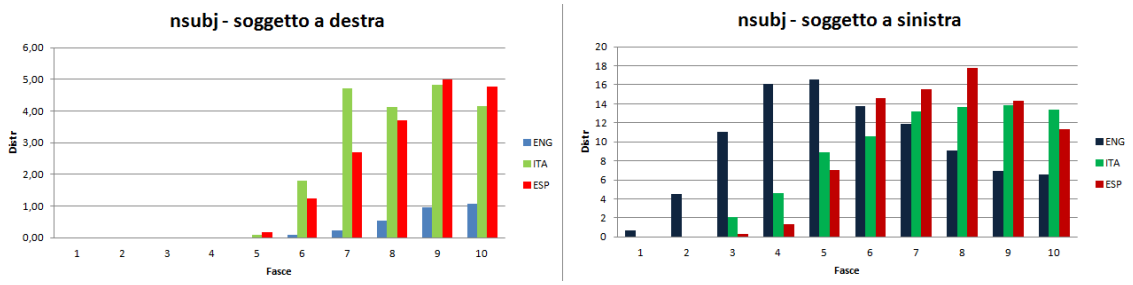


Figura 5.17: Distribuzione (percentuale) delle relazioni *nsubj* sulla base dell'orientamento.

I soggetti che si collocano a sinistra rispetto alla loro testa sono complessivamente i più comuni in tutti e tre i corpora: infatti raggiungono valori percentuali più elevati. In inglese in particolare i primi soggetti che vengono classificati sono costituiti da pronomi personali che appaiono accanto alla testa verbale in frasi principali. Al contrario i soggetti dislocati a destra del predicato sono molto rari, specialmente in inglese, e vengono quindi collocati nelle ultime fasce. Questi ultimi casi sono costituiti da forme interrogative e costruzioni specifiche per ogni lingua. Per l'inglese si tratta di frasi parentetiche (i.e. *"[...], said Bush,[...]"*) e frasi presentative, ovvero rette dalla forma pronominale del verbo essere, che in inglese si realizza come *"there is/there are"*; in questi casi il soggetto è sempre post-verbale. Fra i soggetti post-verbali in italiano e spagnolo troviamo anche frasi alla forma passiva e alcuni enunciati che

collocano il soggetto in una posizione non canonica per scopi enfatici. Si tratta comunque di costruzioni che hanno una frequenza d'uso nella lingua relativamente bassa, e questo giustifica il basso punteggio assegnato da LISCA.

I soggetti in posizione canonica, ovvero preposti al predicato, raggiungono per ogni lingua valori massimi in fasce diverse: italiano e spagnolo presentano un picco di valori in fasce simili (fascia 8 per lo spagnolo e fascia 9 per l'italiano), mentre per quanto riguarda l'inglese il valore più alto si realizza in fascia 5. In quest'ultima lingua, quasi tutte le relazioni soggetto che appaiono entro la fascia 6 hanno come dipendente un pronome personale: questa categoria grammaticale costituisce il 58% dei soggetti totali nell'inglese, mentre in italiano e spagnolo sono rispettivamente il 25% e il 27%. I pronomi personali, indipendentemente dalla lingua, si collocano quasi sempre in posizione canonica, mentre sono i sostantivi ad avere maggiore variabilità. Dal momento che la maggior parte dei soggetti italiani e spagnoli sono nomi, per questi si osserva una distribuzione meno lateralizzata delle relazioni soggetto, tanto che non sembra essere il parametro più significativo nel calcolo della probabilità di questi archi. Ciò che permette a LISCA di discriminare fra una relazione soggetto semplice o complessa è piuttosto la distanza lineare fra testa e dipendente: per primi appaiono i soggetti giustapposti al predicato, e man mano che si avanza nelle fasce la distanza media fra i due cresce. Per italiano e spagnolo, la distanza media fra soggetto e predicato in ultima fascia è di circa 10 parole, in inglese 8.

Interessante notare come le costruzioni con soggetto in posizione non canonica siano quasi totalmente assenti in inglese, mentre in italiano e spagnolo sono maggiormente accettate, seppur raramente. Combinando le informazioni sulla complessità linguistica ottenute dall'analisi di LISCA con la nozione tradizionale di marcatezza [Tusa et al., 2016], possiamo affermare che il soggetto posposto al verbo rappresenta una costruzione marcata, a differenza delle costruzioni con soggetto in posizione preverbale che sono canoniche in tutte e tre le lingue sotto esame. Tuttavia, potendo immaginare di collocare i fenomeni marcati su di una scala graduata, otterremmo che la costruzione con soggetto post-verbale in inglese risulta maggiormente marcata rispetto alla medesima costruzione nelle altre due lingue. Un approccio di questo tipo permette di osservare fenomeni di marcatezza linguistica, solitamente studiati come fenomeno interno alla singola lingua, anche in un contesto multilingua.

Allo stesso modo LISCA permette anche di isolare dipendenze della treebank annotate in modo errato. Potendo prendere in considerazione più parametri contemporaneamente (i.e. lunghezza link e part-of-speech della testa) si possono facilmente

rintracciare le incompatibilità di annotazione. La testa di una relazione soggetto, per esempio, non potrà mai essere un avverbio.

5.4.2 Complemento oggetto diretto

Per la relazione *dobj*, che rappresenta l'oggetto diretto del predicato, non si osservano significative differenze fra lingua e lingua per quanto riguarda la distribuzione; pare che nella risorsa inglese gli oggetti diretti appaiano in quantità superiore, ma ciò non ha ricadute sulla loro struttura sintattica.

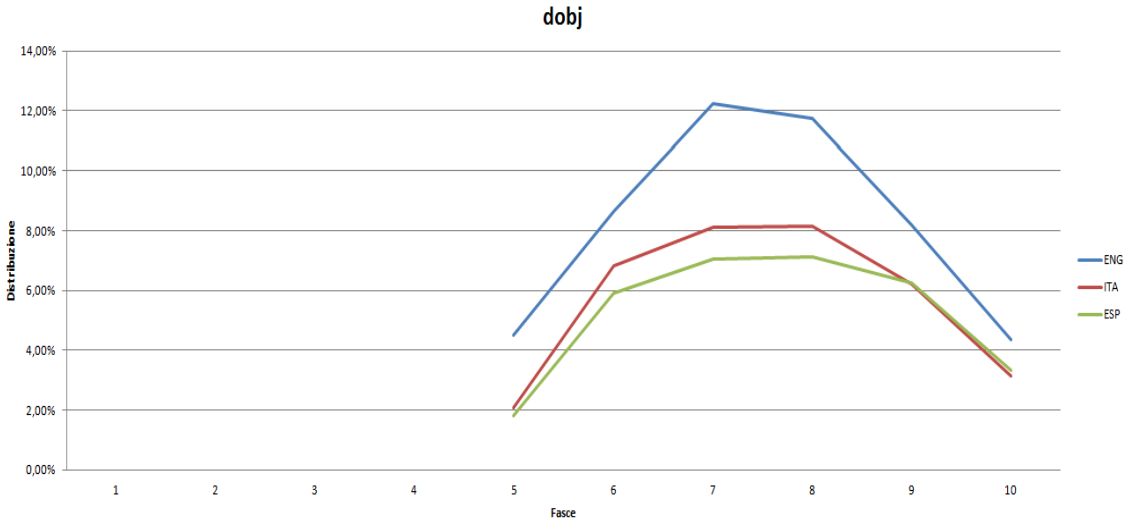


Figura 5.18: Distribuzione (percentuale) delle relazioni *dobj* per fascia.

Trattandosi di lingue ad ordinamento canonico SVO, il complemento oggetto tipicamente si posiziona a destra rispetto alla testa verbale, come dimostra il grafico sull'ordinamento (figura 5.19). In queste costruzioni infatti il comportamento delle tre lingue è molto simile, ad eccezione di minime variazioni poco significative. L'oggetto in posizione pre-verbale è invece ammesso in un numero limitatissimo di casi. In inglese appare prevalentemente come pronome relativo che introduce una subordinata oggettiva (e.g. *"I didn't know that [...]"*) e nelle interrogative (in particolare *wh- questions*).

Per quanto riguarda italiano e spagnolo, i complementi oggetto che precedono il verbo possono essere non solo pronomi relativi, ma anche pronomi clitici (e.g. *"mi"*, *"si"*, *"lo"*, etc.). Tali casi sono piuttosto frequenti in italiano ma non in inglese; tuttavia

questa volta è meglio non parlare di casi di marcatezza visto che in inglese si tratta di costruzioni che vincolano fortemente la posizione dei costituenti, quindi posizionare l'oggetto prima del predicato è obbligatorio e non una scelta comunicativa.

Dall'analisi emerge inoltre che l'oggetto diretto può disporsi anche a grande distanza dalla sua testa nell'ordine lineare della frase e la grande variabilità della lunghezza degli archi è l'elemento che fa sì che LISCA collochi le relazioni *dobj* non prima della fascia 5. La dislocazione dell'oggetto rispetto al predicato è in realtà argomento ampiamente affrontato in letteratura: per esempio è stato osservato che il numero di sillabe che costituiscono l'oggetto ha un forte impatto sulla sua posizione nel caso in cui esso dipenda da un verbo frasale (*Heaviness Effect*) [Merlo, 2017].

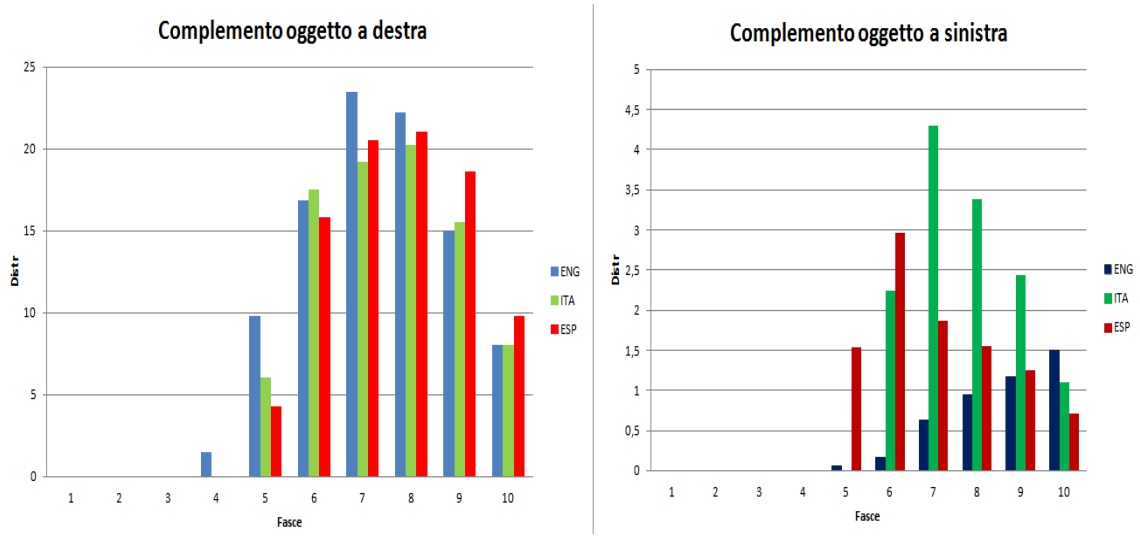


Figura 5.19: Distribuzione (percentuale) delle relazioni *dobj* in base all'orientamento

5.4.3 Aggettivo

L'aggettivo è una parte del discorso variabile che esprime gli attributi relativi alla persona o alla cosa indicata dal sostantivo a cui è riferito e a cui si accorda in genere, numero e caso. Si tenga presente che, come già detto, la treebank inglese classifica gli aggettivi possessivi come pronomi.

L'andamento per le tre lingue nelle prime fasce è piuttosto diverso, ma le tendenze sono uguali. L'aumento di frequenza che tutte e tre le lingue presentano nelle ultime fasce è indice del fatto che l'algoritmo fatica a ritrovare dei comportamenti consistenti per questa classe e assegna bassi punteggi di probabilità ad un gran numero di archi.

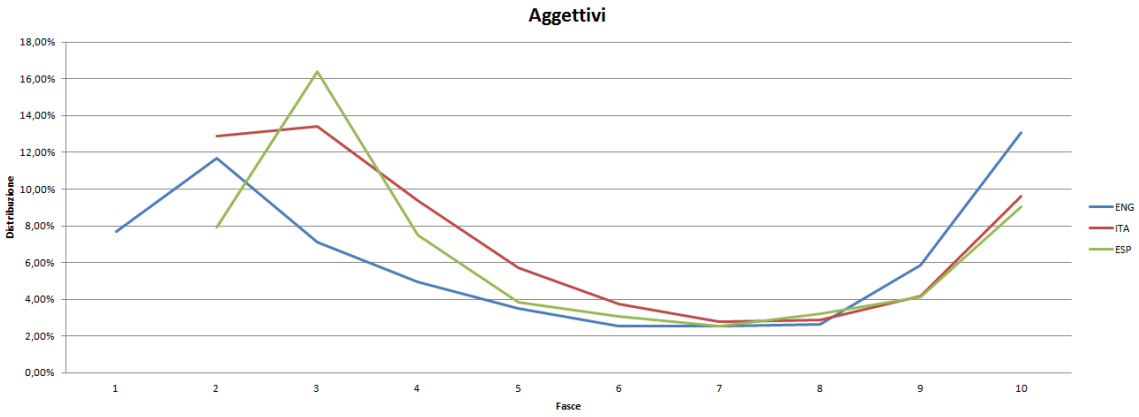


Figura 5.20: Distribuzione (percentuale) di elementi ADJ per fascia.

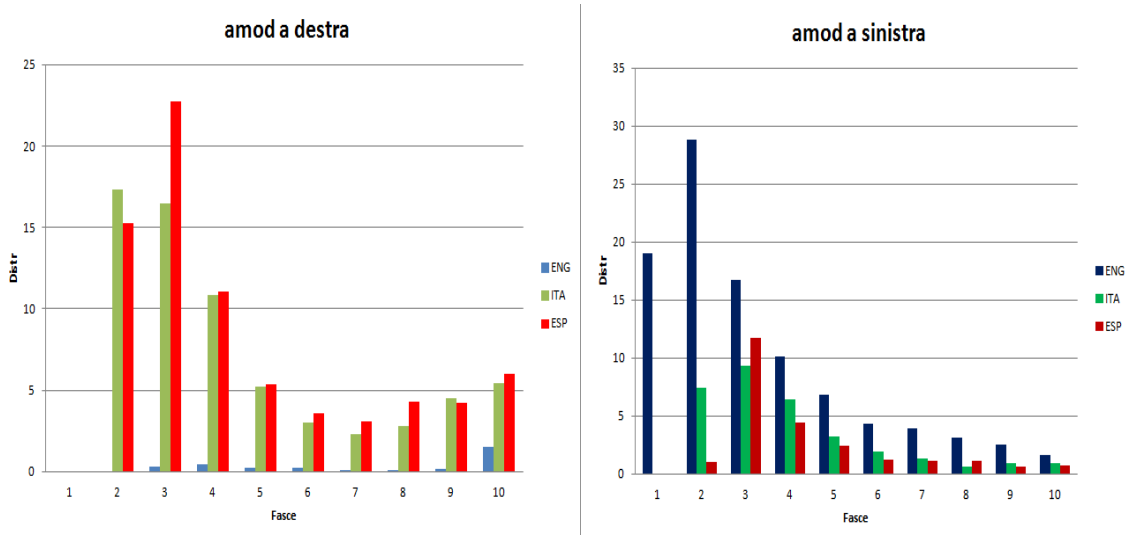


Figura 5.21: Distribuzione (percentuale) delle relazioni *amod* in base all'orientamento.

La relazione sintattica delle Universal Dependencies per i modificatori aggettivali è *amod*. La posizione canonica dell'aggettivo varia nelle lingue: in inglese l'aggettivo si posiziona prima del sostantivo (es. *"the red car"*), mentre accade il contrario nelle altre due lingue (es. *"la macchina rossa"*, *"el coche rojo"*). In italiano e spagnolo è comunque ammessa anche la dislocazione posposta dell'aggettivo, mentre in inglese l'unica posizione ammessa è quella canonica. I link che rappresentano gli aggettivi in posizione prototipica sono relazioni semplici per LISCA, che infatti assegna ad esse punteggi più alti; i comportamenti più anomali che riscontra l'algoritmo riguardano la distanza lineare fra modificatore e testa: nelle ultime fasce si trovano gli aggettivi più distanti dal sostantivo a cui fanno riferimento. Si noti tuttavia che in nessun

caso l'aggettivo si colloca troppo distante rispetto al proprio sostantivo: anche in fascia 10 la lunghezza lineare media per la relazione è di poco inferiore a 3.

La distanza media delle relazioni *amod* (figura 5.22) mette anche in luce un altro fattore: la lunghezza media dei link inglesi è maggiore rispetto a quelle delle altre due lingue. Osservando il contenuto delle fasce si nota come l'inglese faccia maggiore uso della coordinazione di aggettivi rispetto a italiano e spagnolo. In queste ultime due la maggior parte dei link in ogni fascia ha lunghezza 1 o -1; nell'inglese invece, essendoci molti casi di aggettivi coordinati che si riferiscono al solito nome, la distanza lineare fra sostantivo e modificatore aumenta. Per rendere più chiaro il dato si riporta un esempio corredato dei relativi valori. Nell'estratto "*nice and confortable library*", l'arco *amod* si realizza fra "*nice*" (ADJ) e "*library*" (NOUN), con una distanza lineare pari a 3, mentre il secondo aggettivo "*confortable*" ha una relazione *conj* con testa "*nice*".

Si può quindi immaginare di usare LISCA anche per rintracciare scelte di stile che una lingua adotta, utili in particolare nell'apprendimento di una lingua straniera.

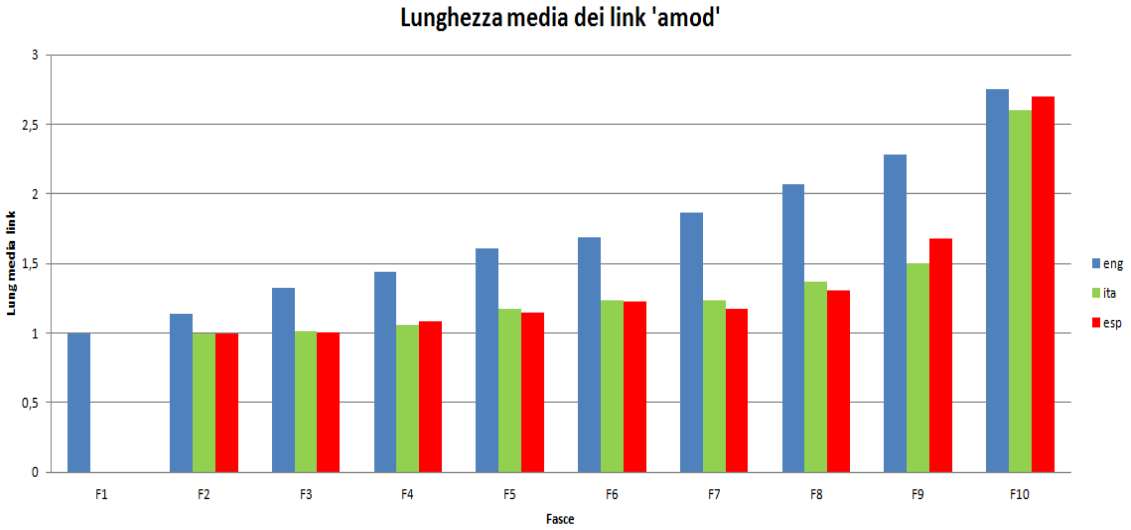


Figura 5.22: Lunghezza media (in parole) fra testa e dipendente negli archi *amod*.

Anche nel caso degli aggettivi le lingue romanze dimostrano di essere molto più flessibili rispetto all'inglese, soprattutto per quando riguarda la dislocazione dell'aggettivo rispetto alla propria testa. In inglese l'aggettivo post-posto si trova solo nel caso dei pronomi indefiniti (e.g. "*something*", "*someone*", "*someplace*", etc.), che lo richiedono necessariamente, mentre in italiano e spagnolo si tratta solitamente di

una scelta comunicativa o stilistica. Per tutte e tre le lingue però la lunghezza media di questi link resta molto bassa e si aggira intorno a 2: la posizione meno standard sembra imporre una maggiore vicinanza fra gli elementi al fine di evitare ambiguità.

Un dato curioso riguarda il modo in cui sono stati ordinati gli aggettivi italiani e spagnoli: in spagnolo nelle prime fasce si trovano tutti e soli i casi di aggettivi numerali, mentre in italiano si osservano solo aggettivi qualificativi. Dal momento che quella in uso è una versione delessicalizzata dell'algoritmo di LISCA, questa suddivisione potrebbe essere indice di una forte correlazione fra la categoria dell'aggettivo e degli altri elementi della frase e meriterebbe una specifica esplorazione.

5.4.4 Avverbio

Gli avverbi sono una delle classi aperte del lessico. Si tratta di una categoria grammaticale di difficile definizione, poiché presenta caratteristiche comuni a classi molto diverse fra loro. Per esempio sono invariabili e non dotati di flessione come le preposizioni, ma svolgono la funzione di modificatori dei verbi, come i sostantivi e gli aggettivi. Inoltre possono anche combinarsi fra loro per giustapposizione e creare locuzioni avverbiali.

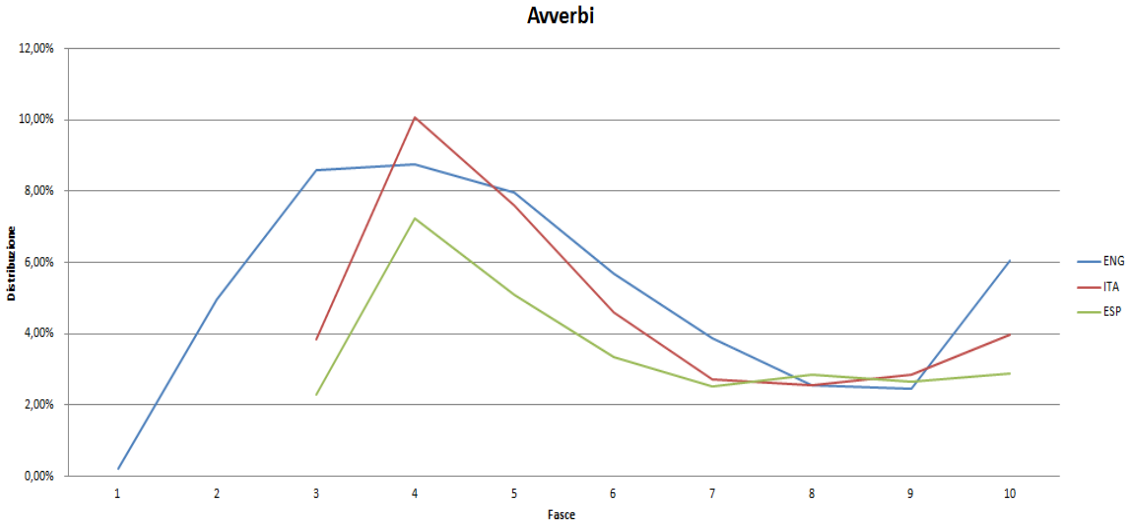


Figura 5.23: Distribuzione (percentuale) degli elementi ADV per fascia.

Sebbene in inglese gli avverbi appaiano più precocemente nelle fasce, complessivamente il comportamento è piuttosto simile ad italiano e spagnolo. È andando ad

osservare i contenuti delle singole fasce che emergono quali siano le differenze che le distinguono. Gli avverbi dell'inglese che ottengono i punteggi di LISCA più alti sono le locuzioni avverbiali (e.g. *"as well"*, *"at least"*, etc.); immediatamente dopo compaiono gli avverbi giustapposti al verbo, nello specifico molti casi di verbi frasali (e.g. *"look down"*, *"go after"*, ...). In italiano e spagnolo troviamo invece gli avverbi giustapposti alla testa verbale nelle prime fasce, mentre le locuzioni avverbiali appaiono successivamente.

Man mano che si procede nell'ordinamento di LISCA, la distanza media fra l'avverbio e la sua testa tende ad aumentare. Nelle fasce intermedie e finali infatti si dispongono i modificatori avverbiali che si collocano distanti dall'elemento cui si appoggiano. Spesso si tratta di relazioni annidate all'interno di frasi subordinate, che ottengono bassi punteggi non tanto per un'anomalia della relazione di modificazione avverbiale quanto piuttosto a causa dell'alto indice di complessità della frase che le ospita.

Nell'inglese tuttavia gli avverbi subiscono un'impennata di valori nell'ultima fascia: si tratta per lo più di avverbi in posizione post-verbale. Queste costruzioni sono molto rare, hanno quindi bassa rilevanza statistica, e si realizzano per casi particolari in cui la lingua inglese consente agli avverbi di assumere posizioni insolite all'interno della frase per scopi enfatici. In mezzo a questi fenomeni di marcatezza tuttavia si nascondono anche diverse annotazioni errate che LISCA intercetta.

5.4.5 Proposizioni subordinate

Le frasi subordinate, dette anche secondarie, sono proposizioni che dipendono logicamente e grammaticalmente da un'altra frase, principale o subordinata a sua volta. In UD sono rappresentate dalle relazioni *acl, acl:relcl, advmod, ccomp e xcomp.*

Si tratta di dipendenze complesse, tanto che LISCA le dispone tutte nelle fasce dalla 6 alla 10. Le frasi secondarie inoltre si collocano solitamente dopo la loro frase reggente. Questa tendenza è molto probabilmente data dal fatto che le subordinate posposte alla principale sono più facili da processare; infatti, quando questa posizione non è rispettata, Pieri et al. [2016] hanno registrato una tendenza alla semplificazione della subordinata stessa sia in termine di numero di parole che la compongono, sia in termini strutturali per quanto riguarda la profondità che raggiunge nel sottoalbero sintattico.

Anche nel caso di annotatori umani queste sono le relazioni su cui si riscontrano i maggiori disaccordi. Non si tratta in questo caso tanto di problemi legati a specifiche discordanti, quanto piuttosto a vere e proprie ambiguità della lingua.

5.4.5.1 *Clausal modifier of noun* e *relative clause modifier* (acl e acl:relcl)

La relazione *acl* viene usata per le proposizioni subordinate che modificano un sostantivo o un pronome. Viene anche usata in caso di proposizioni infinitive introdotte da "il fatto che" (*"the fact that"* in inglese; in spagnolo non si attestano casi simili), e proposizioni infinitive che dipendono da un nome (e.g. *"il caso di pensare"*).

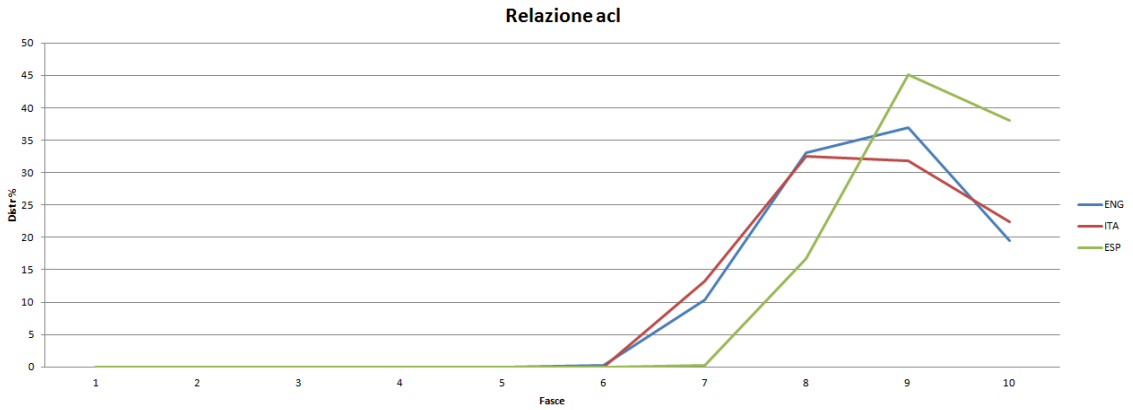


Figura 5.24: Distribuzione (percentuale) delle relazioni *acl* per fascia.

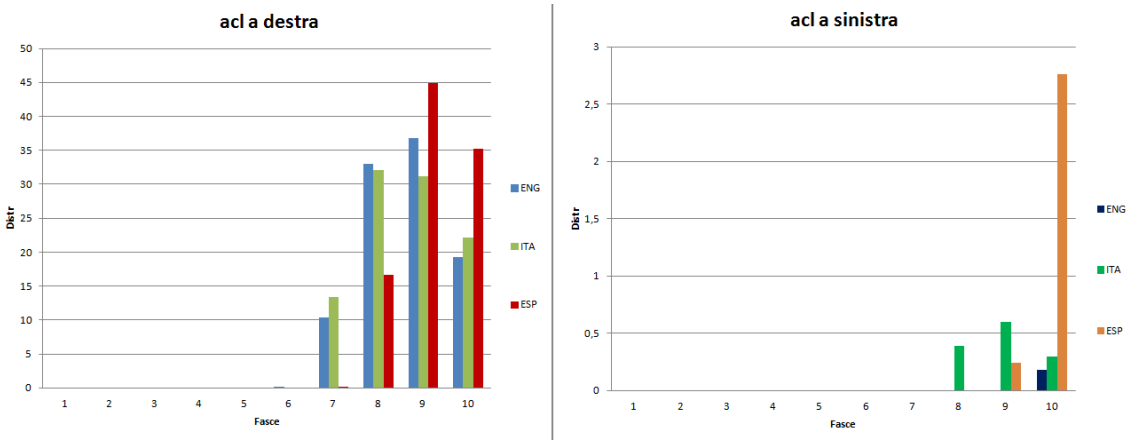


Figura 5.25: Distribuzione (percentuale) delle relazioni *acl* in base all'orientamento.

Nonostante le linee guida di UD circoscrivano i casi d'uso di questa relazione, in spagnolo viene spesso usata per rappresentare, oltre ai casi previsti, anche complementi predicativi e proposizioni gerundive (e.g. *"sigue haciendo"*) per cui invece sono previste etichette differenti. Questo spiega la maggiore quantità di archi *acl* in questa lingua. Si escludano perciò per il momento i dati dello spagnolo, gonfiati dagli errori di annotazione, e si prendano in considerazione solo inglese ed italiano.

Sebbene *acl* ammetta anche lunghe distanze lineari fra il dipendente e la testa nominale (nella fascia 10 la lunghezza media dell'arco è 5), rispetta regole piuttosto rigide per quanto riguarda l'ordinamento della frase. Nella treebank inglese si osservano solo 3 casi in cui la relazione *acl* si posiziona a sinistra della sua testa, ma sono dovuti ad errori di annotazione. In italiano se ne contano un numero maggiore, ma molti di questi archi collegano sostantivi a participi passati che svolgono la funzione di aggettivi, o di aggettivi veri e propri che per forma coincidono col participio del verbo da cui derivano e a cui è stata assegnata la categoria grammaticale sbagliata (e.g. *"tentato omicidio"*, *"comprovati motivi"*, *"avvenuto raggiungimento"*, etc.). Non mancano comunque anche effettivi casi di relazioni *acl*, riconoscibili dalla lunghezza del link, che deve essere almeno pari a 2. In queste frasi l'anomalia nella dislocazione costituisce un caso di costruzione marcata.

La relazione *acl:relcl* viene usata invece per le proposizioni relative. Viene considerata un caso particolare di *acl* e difatti, come questa, si colloca esclusivamente a destra del predicato della reggente, indipendentemente dalla lingua. Anche per questa relazione la lunghezza media degli archi è pari a 5, più alta rispetto a quella di altre dipendenze.

La grande distanza che si osserva fra subordinata e reggente corrispondente è giustificata dalla natura stessa di tali proposizioni. Le frasi secondarie vengono introdotte da altre parti del discorso e possono essere dotate di un loro soggetto, che si colloca fra testa e dipendente, allontanandoli.

Complessivamente, le differenze interlinguistiche che si osservano per questa relazione non sono particolarmente significative.

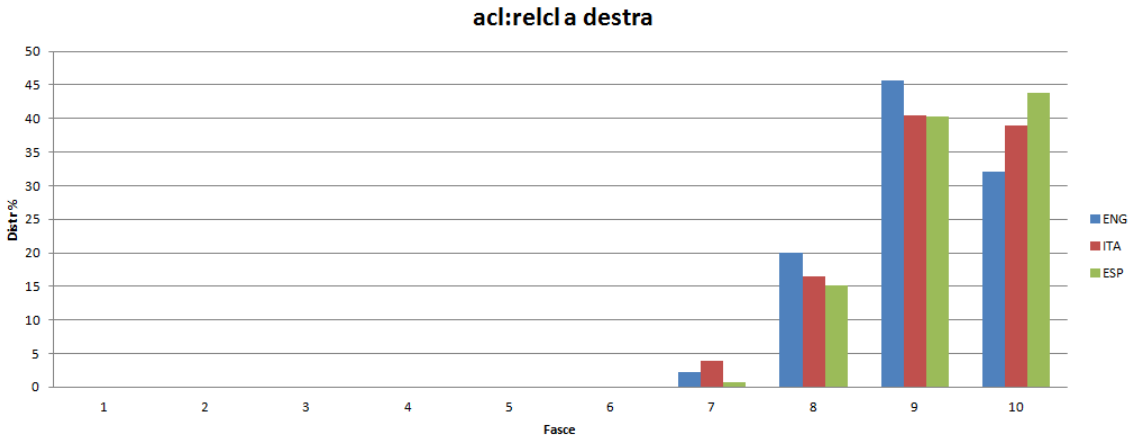


Figura 5.26: Distribuzione (percentuale) delle relazioni *acl:relcl* per fascia. Si posizionano esclusivamente a destra rispetto alla testa.

5.4.5.2 *Adverbial clause modifier* (advcl)

La relazione *advcl* si usa per tutte le proposizioni subordinate introdotte da un avverbio o da una preposizione. Rientrano quindi in questa categoria le proposizioni che la grammatica tradizionale chiama finali, temporali, causali, etc.

Trattandosi di subordinate, anche le proposizioni *advcl* possono trovarsi molto distanti dalla reggente: nelle ultime fasce la lunghezza media dei link si aggira attorno al valore 11 per le tre lingue, che complessivamente non presentano significative discrepanze. Si può notare tuttavia che la libertà di posizione nella disposizione lineare dei costituenti non comporta una libertà anche per quanto riguarda l'ordinamento della frase: queste relazioni si trovano quasi esclusivamente dopo la loro reggente ad esclusione di poche costruzioni marcate, presenti però in tutte e tre le lingue in simile quantità.

La posizione delle subordinate avverbiali rispetto alla principale è fortemente indagata dalla letteratura psicolinguistica, la quale motiva la canonicità dell'ordine lineare posposto sulla base della difficoltà di elaborazione cognitiva della frase (per maggiori dettagli si vedano Diessel [1996, 2008]). Diessel [2001] affronta la questione anche in una prospettiva di studio tipologico e riscontra che il tipo linguistico effettivamente incide sulla posizione della subordinata nell'ordine lineare della frase. Potrebbe essere interessante quindi confrontare i risultati ottenuti qui con altri relativi ad altre lingue.

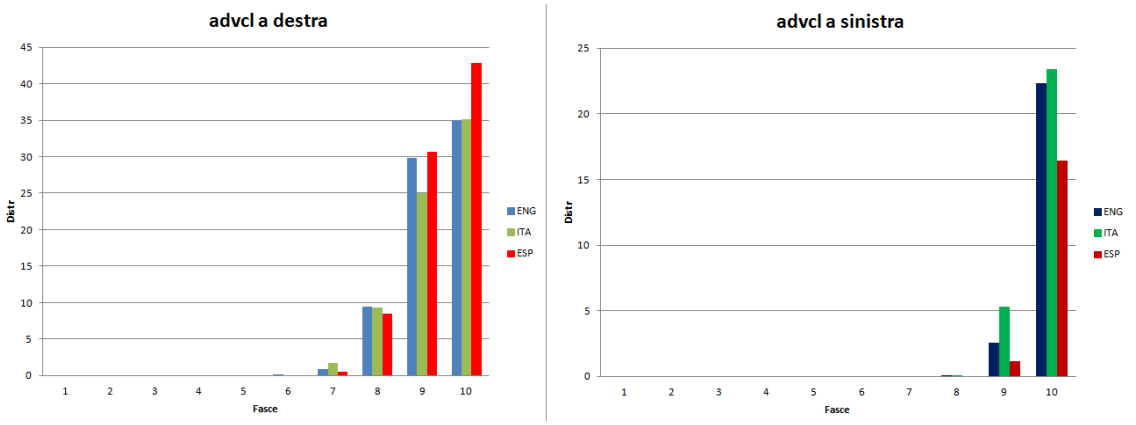


Figura 5.27: Distribuzione (percentuale) delle relazioni *advcl* in base all'orientamento.

5.4.5.3 *Clausal complement* e *open clausal complement* (ccomp e xcomp)

La relazione *ccomp* e la relazione *xcomp* sono spesso di difficile interpretazione anche per gli annotatori umani, difatti si trovano esclusivamente nelle ultime fasce di LISCA.

Le due relazioni servono per rappresentare proposizioni dipendenti da un predicato verbale di cui sono argomento necessario (*core argument*). La differenza fra le due sta nel fatto che nel caso di *ccomp* il soggetto della subordinata non è determinato obbligatoriamente dal verbo reggente (o perché chiaramente espresso o perché determinato anaforicamente o arbitrariamente), mentre nel caso della relazione *xcomp* il soggetto nella subordinata è controllato dalla testa verbale. Questo comporta che spesso i due soggetti coincidano, sebbene tale coincidenza non sia obbligatoria. Si veda come esempio la differenza fra le due frasi seguenti, contenenti entrambe una relazione *xcomp*: "ho promesso di partire" e "ho ordinato di partire". Nella prima frase i due soggetti coincidono (io prometto che io partirò), mentre nella seconda sono diversi (io ho ordinato che qualcun altro parta).

La quasi totalità delle relazioni *ccomp* si colloca a destra rispetto alla propria testa, ovvero la segue. Per tutte e tre le lingue esistono un circoscritto numero di frasi in cui invece l'ordine è invertito. In italiano si tratta di un numero estremamente ristretto di casi (21), mentre in inglese e spagnolo sono più numerosi (116 in spagnolo e 117 in inglese).

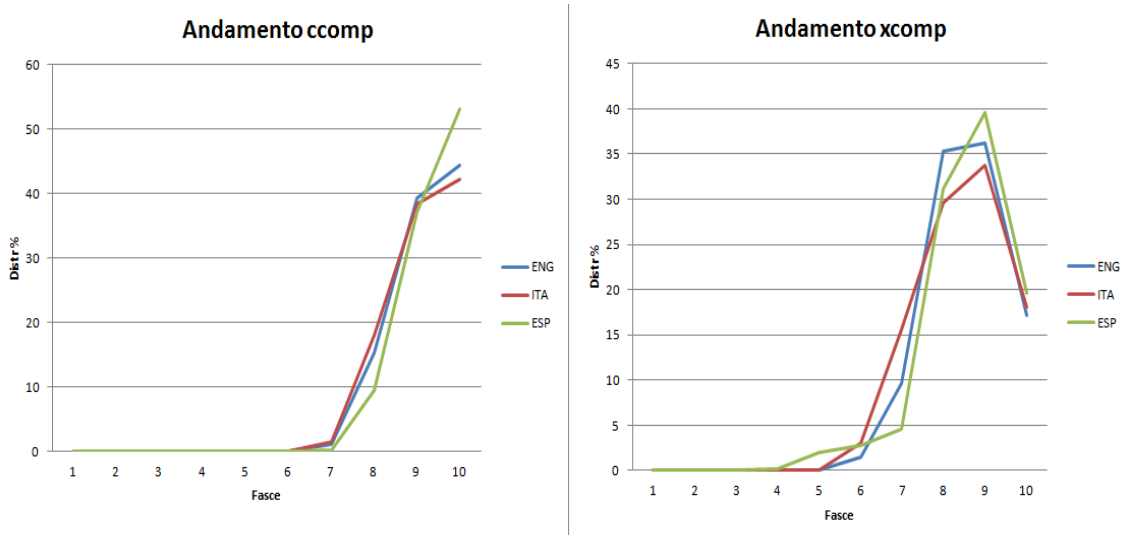


Figura 5.28: Distribuzione (percentuale) delle relazioni *ccomp* e *xcom* per fascia.

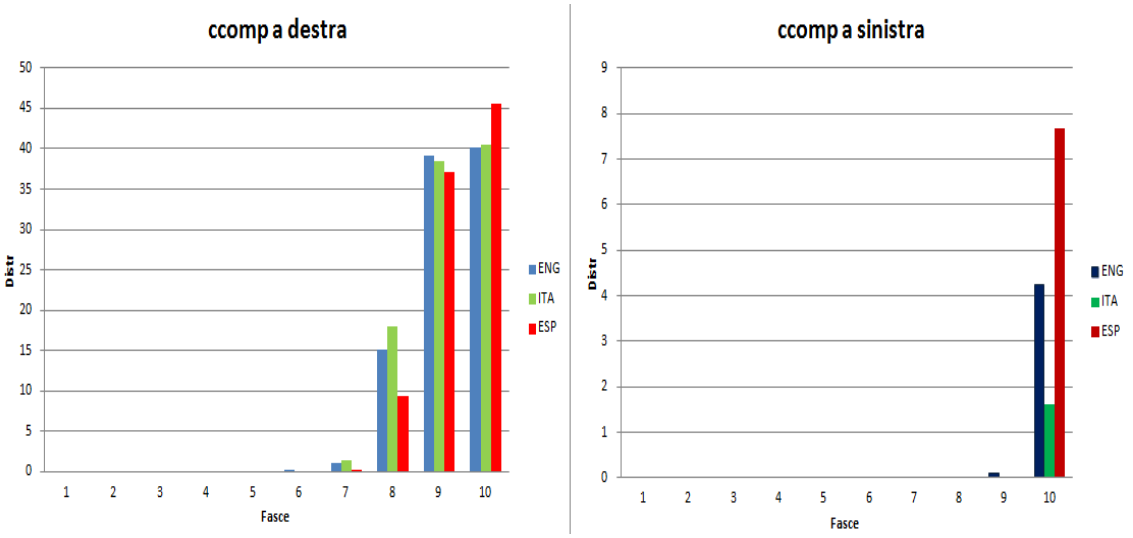


Figura 5.29: Distribuzione (percentuale) della relazione *ccomp* in base all'orientamento.

Per quanto riguarda l'italiano si tratta in effetti di costruzioni che presentano un ordine marcato per scopi enfatici (e.g. *"[...] a preoccupare la Croazia è [...]"*). Le linee guida per inglese e spagnolo per quanto riguarda questa relazione invece sono meno dettagliate rispetto a quelle dell'italiano, per cui si osserva una disomogeneità nei criteri di annotazione anche all'interno della stessa lingua.

In inglese e spagnolo per esempio la relazione *ccomp* preposta al verbo viene usata nelle frasi in cui è presente un discorso riportato indirettamente, ovvero introdotto da

un verbo dichiarativo (e.g. *"John said"*, *"I assure you"*, *"declara"*, *"dije"*, etc.). Per questi casi in realtà le UD suggerirebbero di usare la relazione *parataxis*, che serve proprio a marcare i casi in cui fra due proposizioni non sussiste una chiara relazione di subordinazione o coordinazione ma piuttosto una giustapposizione che le rende potenzialmente indipendenti l'una dall'altra. Sebbene l'uso di *ccomp* in questi casi non si possa considerare un vero e proprio errore di annotazione, si tratta comunque di una disomogeneità significativa fra treebank.

Trattandosi di relazioni subordinate, come già detto la lunghezza media dei link fra testa e dipendente è piuttosto elevata e si aggira intorno al valore 4 per italiano e inglese. Più alta invece per lo spagnolo che ha una lunghezza media pari a 5, influenzata dal fatto che in ultima fascia si collocano archi con distanze anche molto elevate.

Proprio in virtù della sua struttura che la lega fortemente alla propria reggente, le proposizioni *xcomp* si trovano invece sempre posposte e mai troppo distanti dalla loro testa. La lunghezza media dei link infatti è circa 2 per tutte e tre le lingue. Fra queste però ricadono anche i complementi predicativi del soggetto costituiti da aggettivi. Questi sono particolarmente frequenti in inglese (e.g. *"goes crazy"*, *"seems prudent"*) che difatti ne possiede un numero maggiore di esempi all'interno della sua risorsa .

Complessivamente quello che si osserva per queste due relazioni di dipendenza è che le tipologie di casi che rappresentano sono piuttosto disomogenei fra lingue e che quindi meritino un'indagine più approfondita per rendere le loro linee guida e le loro applicazioni più universali e comparabili.

5.5 Riepilogo

I risultati ottenuti dalla ricerca si possono riassumere all'interno delle categorie di seguito riportate. Le prime tre riguardano gli aspetti linguistici che sono emersi, mentre le ultime due descrivono le applicazioni pratiche dell'algoritmo.

Casi di marcatezza tipologica: La metodologia di indagine messa a punto ha permesso di rintracciare casi di marcatezza tipologica all'interno delle treebank.

Gli andamenti per fascia delle classi grammaticali e delle relazioni di dipendenza di italiano e spagnolo spesso si sono rivelati praticamente sovrapponibili (si vedano i grafici nelle figure 5.6, 5.7, 5.8, 5.16, 5.18, 5.24 e 5.28 per i casi più evidenti). Questo dimostra che LISCA consente di intercettare le costruzioni e i comportamenti vincolati dal tipo linguistico. In aggiunta a ciò, l'osservazione dei contenuti delle singole fasce ha dimostrato che l'ordinamento sulla base della plausibilità effettivamente circoscrive le costruzioni marcate fra quelle che ottengono i punteggi più bassi. È stato quindi dimostrato che la nozione di marcatezza, intesa come strutture che deviano dalla canonicità della lingua e come rarità nei testi, può essere efficacemente usata nell'interpretazione dei risultati per riconoscere i fenomeni di marcatezza di una treebank grazie a LISCA.

Dall'intersezione fra queste due informazioni e dall'applicazione dell'algoritmo in un contesto multilingua, è stato possibile individuare i casi di marcatezza tipologica, ovvero quelle costruzioni che sono marcate in tutte le lingue della stessa tipologia. È questo il caso dei soggetti post-verbali, marcati in tutte e tre le lingue, sebbene con gradi diversi, le quali difatti appartengono alla stessa tipologia sintattica. Questa metodologia di indagine si è rivelata efficace e quindi si immagina che l'aggiunta di altre lingue al confronto, del medesimo o diverso tipo, permetterebbe di delineare un quadro più dettagliato delle caratteristiche di ciascun gruppo osservato.

Costruzioni sintattiche specifiche per una determinata lingua: i risultati di questa ricerca hanno fatto emergere numerose informazioni linguistiche su comportamenti specifici di ciascuna delle tre lingue, mettendo in primo piano le similarità e le differenze più significative che le contraddistinguono.

In alcuni casi le discrepanze osservate erano legate alla rappresentazione sintattica utilizzata (più nel dettaglio in seguito), ma molti di essi invece rispecchiavano vere e proprie caratteristiche specifiche della lingua. Si tratta per esempio dell'etichetta *compound*, molto frequente in inglese ma non nelle altre lingue proprio perché rispecchia il principio generativo che adotta l'inglese per creare parole nuove.

Scelte d'uso della lingua: lo studio degli aggettivi (intesi sia come POS ADJ che come relazione *amod*) ha messo in luce che LISCA può essere usato anche per intercettare scelte stilistiche di una lingua, come la tendenza della lingua inglese ad usare questi modificatori in coppia piuttosto che in isolamento.

Si tratta in questo caso di un tipo di risultato del tutto inatteso, le cui potenzialità andrebbero esplorate nel dettaglio in futuro.

Errori di annotazione: fra gli archi che si collocano più in basso dell'ordinamento di LISCA si riscontra la presenza di un numero cospicuo di archi a cui è stato assegnata un'etichetta errata.

Sebbene LISCA fosse già stato efficacemente adoperato nella ricerca di errori di annotazione, sono comunque emersi anche casi in cui l'errore non riguarda direttamente l'arco quanto piuttosto altre relazioni della frase in cui l'arco in oggetto è inserito.

Questi casi servono a dimostrare che l'algoritmo tiene in considerazione parametri linguistici e strutturali e non si limita al calcolo delle occorrenze di un arco all'interno di una risorsa annotata.

Inconsistenze nei criteri di annotazione: dal confronto fra più risorse annotate è stato possibile osservare il comportamento di singole relazioni e classi grammaticali per verificare se il loro comportamento subisse variazioni di lingua in lingua. Questo metodo ha permesso di rintracciare anche i casi in cui i comportamenti divergevano a causa dei diversi criteri di annotazione adottati: i più eclatanti riguardano i pronomi nella risorsa inglese (grafico in figura 5.8) e gli oggetti indiretti in quella spagnola.

Tali inconsistenze devono essere tenute in considerazione nella realizzazione degli studi multilingua perché potrebbero dare spiegazione ad alcune delle discrepanze che si osservano. In uno studio di McDonald et al. [2011] sono stati usati dei modelli delessicalizzati di alcune lingue sorgente per addestrare parser che realizzassero l'analisi su una serie di lingue target, diverse dalle prime. Lo studio ha sorprendentemente dimostrato che la vicinanza tipologica fra lingua sorgente e target non aveva necessariamente un impatto positivo sul parsing: il danese, usato come modello di annotazione per lo svedese, ha ottenuto bassissimi punteggi di accuratezza. Ciò si pensa dovuto ai diversi criteri di annotazione usati dalle due lingue, che causano delle inconsistenze anche per costruzioni molto simili. Lo studio presentato tuttavia si limitava ad attestare la distanza fra treebank, ipotizzando delle discrepanze nell'annotazione di alcune costruzioni. Il metodo qui utilizzato in una prospettiva di indagine simile permetterebbe di entrare nel dettaglio rintracciando gli elementi grammaticali e i costrutti specifici che sono causa delle divergenze più significative.

A quelli citati si affiancano anche casi, come quello di *xcomp* e *ccomp*, per cui si osserva una certa inconsistenza di annotazione anche all'interno della stessa risorsa, dovuta o alla mancanza di linee giuda chiare o alla difficoltà propria della relazione da assegnare.

Capitolo 6

Conclusioni

Il presente elaborato ha descritto metodologie, strumenti e risultati di un'indagine tipologica svolta su tre lingue, due tipologicamente molto simili e una più distante, adottando un approccio linguistico-computazionale con l'obiettivo di far emergere similarità e differenze fra le lingue stesse. La ricerca è stata svolta grazie all'utilizzo delle risorse linguistiche annotate manualmente o semi-automaticamente del progetto Universal Dependencies. Gli strumenti che hanno permesso di ottenere i dati oggetto della discussione sono la catena di annotazione linguistica UDPipe, con cui tre corpora monolingue sono stati annotati automaticamente secondo lo standard UD, e LISCA, un software che assegna ad ogni arco di un corpus annotato, in questo caso gold, un valore di plausibilità e che crea un ordinamento degli archi sulla base di tale punteggio. Come si è detto, in questo contesto per plausibilità si intende la probabilità che un arco si realizzi nella treebank, ovvero che una relazione grammaticale si realizzi nel linguaggio. Grazie a tali punteggi è stato possibile confrontarle e inferire dei giudizi sul comportamento delle tre lingue.

Sebbene l'ipotesi di ricerca iniziale ambisse a rintracciare e confrontare le costruzioni all'interno delle treebank sulla base dei punteggi di plausibilità di LISCA, di fatto la metodologia usata ha permesso di raggiungere dei risultati utili per ricavare informazioni molto più ricche ed interessanti. L'approccio, che ha sfruttato i dati prodotti da uno strumento nato per essere usato in un ambito fortemente applicativo, si è rivelato un metodo efficace per lo studio delle lingue da un punto di vista tipologico.

Sul versante dello studio linguistico, il metodo linguistico-computazionale e lo schema di annotazione UD si sono dimostrati efficaci nel far emergere le caratteristiche

che accomunano o distinguono le lingue sulla base della tipologia. In questo senso uno dei casi più significativi fra quelli osservati riguarda la dislocazione pre- o post-nominale degli aggettivi. Inoltre, dal confronto inter-linguistico delle singole fasce di LISCA, è stato possibile derivare alcuni principi universali relativi alla complessità: come è stato discusso, infatti, il punteggio di LISCA può essere interpretato anche come un indice di complessità linguistica. Quelle costruzioni che hanno ottenuto alti punteggi in tutte e tre le lingue sotto esame rappresentano fenomeni linguistici che si possono considerare universalmente semplici; viceversa, nel caso di punteggi bassi, sono stati considerati universalmente complessi. Questi casi rispecchiano una complessità effettiva legata alla costruzione, indipendente dalla lingua in cui essa appare. Le costruzioni che invece risultano complesse (o semplici) solo in un sottoinsieme di lingue, subiscono l'influenza della tipologia di appartenenza della lingua in questione; in questo caso si è parlato piuttosto di universali di tendenza. Partendo dal dato di LISCA, si è affrontato allo stesso modo anche lo studio dei fenomeni di marcatezza tipologica. Un tale approccio alla marcatezza, che non solo si avvicina al fenomeno in una prospettiva multilingua ma anche che si basa sul valore di plausibilità piuttosto che su una semplice distribuzione di frequenza, è assolutamente innovativo.

Il presente studio è nuovo anche nel filone degli studi tipologici. Il progetto UD, per sua stessa natura, crea i presupposti per la realizzazione di ricerche multilingua, e infatti i tipologi si sono dimostrati sin da subito molto interessati all'iniziativa e ai criteri di annotazione definiti, riconoscendo in essi nuove possibilità di ricerca. Tuttavia fino a questo momento il loro contributo è stato prettamente teorico, con la ridefinizione di criteri e relazioni sintattiche in modo da rendere le risorse sempre più confrontabili. Le uniche linguiste che per ora si sono avvicinate alle treebank UD con fini di indagine tipologica sono le già citate Gulordava e Merlo, le quali tuttavia si sono concentrate solo sulla relazione aggettivo. Grazie a questa tesi invece è stato dimostrato chiaramente che da un confronto multilingua che coinvolge l'intero tagset possono emergere numerose informazioni, alcune delle quali inaspettate, come ad esempio le scelte stilistiche delle lingue.

Infatti, nonostante le differenze morfosintattiche fra tipi linguistici siano argomento di ricerca da lungo tempo, grazie a questa metodologia paiono emergere nuove differenze, eventualmente legate ad usi o registri specifici. Potrebbe dimostrarsi ancora più interessante in questo senso l'uso di corpora contenenti lo stesso testo tradotto in lingue differenti. Come sostiene Umberto Eco, tradurre significa "dire quasi la stessa cosa": l'opera di traduzione è una manipolazione del testo che fa sì che un

medesimo contenuto venga espresso in lingue diverse rispettando la grammaticalità di ciascuna di esse. Un confronto come quello condotto qui permetterebbe di individuare quali costruzioni sono condivise da più lingue e quali invece non hanno nessun corrispondente, e richiedono dunque una trasformazione più consistente.

Contemporaneamente questa ricerca dimostra l'utilità dello strumento LISCA anche per quanto riguarda aspetti applicativi legati al parsing automatico e alla revisione manuale (o semi-manuale) delle risorse linguistiche, nonché nell'individuazione delle relazioni di dipendenza più critiche che richiedono delle linee guida più dettagliate per il loro utilizzo.

Pur avendo svolto l'indagine su risorse *gold* annotate manualmente, l'algoritmo ha permesso di rintracciare gli archi la cui annotazione presentava delle incorrettezze e, allo stesso tempo, dal confronto parallelo fra treebank ha fatto emergere casi specifici di costruzioni in cui lo schema di annotazione UD non viene applicato in maniera univoca da tutte le lingue. I casi più eclatanti di quest'ultimo tipo riguardano i pronomi nella risorsa inglese e gli oggetti indiretti in spagnolo. Estrapolare da una treebank annotata gli archi che hanno un basso indice di plausibilità invece può essere usato per facilitare il lavoro dei revisori, i quali possono focalizzare la loro attenzione solo sui casi che veramente lo necessitano rendendo molto più rapide ed efficienti le operazioni di controllo.

Infine, si vuole fare qui un'ulteriore riflessione sulle potenzialità di questa metodologia applicata alle treebank UD. Dal momento che lo schema di annotazione è pensato per essere valido per qualsiasi lingua, indipendentemente dal tipo, anche la valutazione delle performance di un parser dovrebbe tenere conto di questo fattore e LISCA può rivelarsi utile proprio nella definizione di una nuova metrica di valutazione più adeguata a soddisfare tali esigenze.

Si pensi alla differenza fra lingue sintetiche e lingue analitiche: le prime, di cui fa parte per esempio il finlandese, sono morfologicamente codificate per quanto riguarda la componente sintattica, mentre le seconde veicolano le stesse informazioni per mezzo di elementi funzionali come preposizioni ed articoli. Le categorie grammaticali funzionali, come si è visto, sono le più semplici da assegnare; per questa ragione un parser valutato in termini di LAS e UAS, metriche che non tengono conto della distinzioni fra classi grammaticali aperte e chiuse, otterrà punteggi più alti nel caso di lingue analitiche piuttosto che di lingue sintetiche. I risultati di LISCA sembrano allinearsi con la posizione presa da Nivre [2016]: l'autore infatti sostiene che il rin-

novato ambito del parsing multilingua, favorito dalla nascita di UD, necessita di una revisione anche per quanto riguarda le metriche di valutazione delle performance, le quali dovrebbero tenere conto della complessità relativa alla POS o alla relazione di dipendenza e attribuire il giusto peso ad ogni assegnamento. Grazie ai dati che sono stati ottenuti qui è già possibile giustificare una suddivisione del tagset fra relazioni che meritano di essere oggetto di valutazione e relazioni poco informative in questo senso; inoltre il punteggio di plausibilità può essere usato come parametro per stabilire il peso di ogni relazione di dipendenza. Si può dunque affermare che LISCA, o uno strumento ad esso equivalente, possono essere adoperati per stabilire l'efficacia di uno strumento di parsing a dipendenze.

Sulla base dei risultati ottenuti e delle informazioni che ne sono state ricavate, si può affermare che attraverso questa ricerca è stato messo a punto un metodo di indagine per i tipi linguistici innovativo ed efficace, che apre la strada a nuove prospettive di ricerca ancora inesplorate che sembrano però estremamente promettenti. Si potrebbe per esempio come primo passo realizzare il medesimo studio su un numero maggiore di lingue, così da ottenere confronti più accurati sia per singole relazioni che per comportamenti generici. Inoltre non bisogna scordare le possibili applicazioni pratiche di questo metodo nel miglioramento delle risorse linguistiche e nella valutazione degli schemi di annotazione e degli strumenti di parsing. In conclusione si vuol far riflettere sul fatto che i risultati ottenuti dimostrano chiaramente che le potenzialità degli strumenti e della metodologia qui adottata sono vastissime, sebbene informazioni altrettanto ricche siano ancora nascoste all'interno delle risorse linguistiche. È opportuno dunque che tali risorse continuino ad essere esplorate con metodi sempre nuovi poiché la ricerca di certi fenomeni può facilmente risultare nella scoperta di dati non previsti, come avvenuto in questo caso. Pertanto molto può ancora essere fatto, sia nel rinnovato filone di ricerca multilingua sull'indagine tipologica, sia nello sviluppo degli strumenti per realizzare tali studi.

Bibliografia

Bharat Ram Ambati, Mridul Gupta, Samar Husain, and Dipti Misra Sharma. A High Recall Error Identification Tool for Hindi Treebank Validation. In *LREC*, 2010.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *Proceedings of ACL*, 2016.

Edna Andrews. *Markedness Theory*. Duke University Press, May 1990.

Enrique Henestroza Anguiano and Marie Candito. Parse Correction with Specialized Models for Difficult Attachment Types. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 1222–1233, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Giuseppe Attardi and Massimiliano Ciaramita. Tree Revision Learning for Dependency Parsing. In *HLT-NAACL*, pages 388–395, 2007.

Roger Bacon. *The Greek grammar of Roger Bacon and a fragment of his Hebrew grammar*. The University Press, 1902.

Gaetano Berruto and Massimo Cerruti. *La linguistica. Un corso introduttivo*. UTET Universitá, Novara, May 2011.

Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. Bias and Agreement in Syntactic Annotations. *CoRR*, 2016.

Balthasar Bickel. Typology in the 21st century: major current developments. *Linguistic Typology*, 11(1):239–251, 2007.

Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 89–97. Association for Computational Linguistics, 2010.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Leonardo Lesmo. Building a Treebank for Italian: a Data-driven Annotation Schema. In *roceedings of the Second International Confer ence on Langua ge Resour ces and Evaluation LREC*, pages 99–106, Athens, Greece, 2000.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. Harmonization and merging of two italian dependency treebanks. In *LREC 2012 Workshop on Language Resource Merging Workshop Programme*, page 23, 2012.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, 2013.

Adriane Boyd, Markus Dickinson, and W. Detmar Meurers. On Detecting Errors in Dependency Treebanks. *Research on Language and Computation*, 6(2):113–137, October 2008.

Joan Bresnan. Linguistics: The Garden and the Bush. *Computational Linguistics*, 42(4):599–617, September 2016.

Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics, 2006.

Andrew Carnie. *Syntax: A Generative Introduction*. John Wiley & Sons, April 2013.

John A. Carroll. Statistical parsing. In *Handbook of natural language processing*. CRC Press, 2000.

Wenliang Chen, Zhenghua Li, and Min Zhang. Tutorial: Dependency Parsing: Past Present and Future. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014.

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 1965.

Noam Chomsky and Howard Lasnik. The theory of principles and parameters. *Syntax: An international handbook of contemporary research*, 1:506–569, 1993.

Michael John Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 184–191. Association for Computational Linguistics, 1996.

Bernard Comrie. Language universale and linguistic typology: data-bases and explanations. *STUF-Language Typology and Universals*, 46(1-4):3–14, 1993.

Vivian J. Cook. Chomsky's universal grammar and second-language learning. *Applied Linguistics*, 6:2, 1985.

William Croft. *Typology and universals*. Cambridge University Press, 2002.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. Old School vs. New School: Comparing Transition-Based Parsers with and without Neural Network Enhancement. 2016.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa, 2006.

Felice Dell'Orletta. Ensemble system for Part-of-Speech tagging. *Proceedings of EVALITA*, 9:1–8, 2009.

Felice Dell'Orletta and Giulia Venturi. ULISSE: una strategia di adattamento al dominio per l'annotazione sintattica automatica. In Edoardo Maria Ponti and Marco Budassi, editors, *Compter parler soigner: tra linguistica e intelligenza artificiale: atti: Pavia, Collegio Ghislieri, 15-17 dicembre 2014*, Atti. Pavia University Press, Pavia, prima edizione edition, 2016.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. ULISSE: An Unsupervised Algorithm for Detecting Reliable Dependency Parses. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL 2011, pages 115–124, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Linguistically-driven Selection of Correct Arcs for Dependency Parsing. *Computacion y Sistemas*, 17 (2):125–136, 2013.

Markus Dickinson. Detecting errors in automatically-parsed dependency relations. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 729–738. Association for Computational Linguistics, 2010.

Markus Dickinson and W. Detmar Meurers. Detecting inconsistencies in treebanks. In *Proceedings of TLT*, volume 3, pages 45–56, 2003.

Markus Dickinson and Amber Smith. Detecting dependency parse errors with minimal resources. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 241–252. Association for Computational Linguistics, 2011.

Holger Diessel. Processing factors of pre-and postposed adverbial clauses. In *Annual Meeting of the Berkeley Linguistics Society*, volume 22, pages 71–82, 1996.

Holger Diessel. The Ordering Distribution of Main and Adverbial Clauses: A Typological Study. *Language*, 77(3):433–455, 2001.

Holger Diessel. Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive linguistics*, 19(3):465–490, 2008.

Wolfgang U. Dressler. *Leitmotifs in natural morphology*, volume 10. John Benjamins Publishing, 1987.

Matthew S. Dryer. Frequency and pragmatically unmarked word order. *Word order in discourse*, 30:105, 1995.

Fred R. Eckman. Typological markedness and second language phonology. *Phonology and second language acquisition*, 36:95–115, 2008.

Giuliana Fiorentino. Complessita linguistica e variazione sintattica. *Studi italiani di linguistica teorica ed applicata (SILTA)*, 38(2):281, 2009.

Bice Mortara Garavelli. *Prontuario di punteggiatura*. Gius.Laterza & Figli Spa, September 2014.

Edward Gibson. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76, August 1998.

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202, 2001.

Daniel Gildea and David Temperley. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310, 2010.

Joseph H. Greenberg. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194, 1960.

Joseph H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113, 1963.

Kristina Gulordava and Paola Merlo. Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *CoNLL*, pages 247–257, 2015.

Jeanette K. Gundel, Kathleen Houlihan, and Gerald A. Sanders. Markedness and distribution in phonology and syntax. In *Markedness*, pages 107–138. Springer, 1986.

Keith Hall and Vaclav Novak. Corrective modeling for non-projective dependency parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 42–52. Association for Computational Linguistics, 2005.

M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1 edition edition, May 1976.

Martin Haspelmath. Against markedness (and what to replace it with). *Journal of linguistics*, 42(01):25–70, 2006.

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

David G. Hays. Dependency theory: A formalism and some observations. *Language*, 40(4):511–525, 1964.

Peter Hellwig. Dependency unification grammar. In *Proceedings of the 11th coference on Computational linguistics*, pages 195–198. Association for Computational Linguistics, 1986.

Richard A. Hudson. *Word grammar*. Blackwell Oxford, 1984.

Elizabeth Hume. Deconstructing markedness: A predictability-based approach. In *Proceedings of BLS*, volume 30, pages 182–198. Citeseer, 2004.

Roman Jakobson. Zur Struktur des russischen Verbums. *Travaux du Cercle Linguistique de Prague*, page 240, 1932.

Roman Jakobson. *Child language, aphasia and phonological universals*, volume 72. Walter de Gruyter GmbH & Co KG, 1968.

Daisuke Kawahara and Kiyotaka Uchimoto. Learning Reliability of Parses for Domain Adaptation of Dependency Parsing. In *IJCNLP*, volume 8, 2008.

Sandra Kubler, Ryan McDonald, and Joakim Nivre. Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127, 2009.

Henry Kucera. Markedness and frequency: A computational analysis. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 167–173. Academia Praha, 1982.

Yves Lecerf. Programme des conflits, modele des conflits. *Bulletin bimestriel de l'ATALA*, 1(4):11–18, 1960.

Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. *Testo e computer. Elementi di linguistica computazionale*. Carocci, Roma, May 2005.

Stephen C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, 2000.

Dekang Lin. On the structural complexity of natural language sentences. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 729–733. Association for Computational Linguistics, 1996.

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

Ricardo Mairal and Juana Gil. *Linguistic universals*. Cambridge University Press, 2006.

Christopher Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? *Computational Linguistics and Intelligent Text Processing*, pages 171–189, 2011.

Solomon Marcus. Sur la notion de projectivite. *Mathematical Logic Quarterly*, 11 (2):181–192, 1965.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source Transfer of Delexicalized Dependency Parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 62–72, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, and others. Universal Dependency Annotation for Multilingual Parsing. In *ACL (2)*, pages 92–97, 2013.

Igor Mel'cuk. Dependency syntax. *State University of New York Press, Albany, NY*, 1988.

Paola Merlo. Some Recent Results on Cross-Linguistic, Corpus-Based Quantitative Modelling of Word Order and Aspect. In *Formal Models in the Study of Language*, pages 451–464. Springer, 2017.

Bettina Mohr, Freidemann Pulvermuller, and Eran Zaidel. Lexical decision after left, right and bilateral presentation of function words, content words and non-words: Evidence for interhemispheric interaction. *Neuropsychologia*, 32(1):105–124, January 1994.

Simonetta Montemagni. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, 42(1): 145–172, 2013.

Simonetta Montemagni and Maria Simi. The Italian dependency annotated corpus developed for the CoNLL-2007 shared task. Technical report, Technical report, ILC-CNR, 2007.

Edith A. Moravcsik. *Introducing Language Typology*. Cambridge University Press, 2013.

Joakim Nivre. Dependency grammar and dependency parsing. *MSI report*, 5133 (1959):1–32, 2005.

Joakim Nivre. Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16. Springer, 2015.

Joakim Nivre. Universal dependency evaluation. *Unpublished paper*, 2016.

Joakim Nivre and Jens Nilsson. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 99–106. Association for Computational Linguistics, 2005.

Joakim Nivre, Johan Hall, Sandra Kubler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932, 2007a.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gulsen Eryigit, Sandra Kubler, Svetoslav Marinov, and Erwin Marsi. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13 (02):95–135, 2007b.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and others. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, 2016.

Doris L. Payne. *The Pragmatics of Word Order: Typological Dimensions of Verb Initial Languages*. De Gruyter Mouton, Berlin, reprint 2013 ed. edition edition, August 1990.

Slav Petrov. Announcing syntaxnet: The world's most accurate parser goes open source. *Google Research Blog, May*, 12:2016, 2016.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. *arXiv:1104.2086 [cs]*, April 2011.

Giulia Pieri, Dominique Brunato, and Felice Dell'Orletta. Studio sull'ordine dei costituenti nel confronto tra generi e complessita. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, December 2016.

Barbara Plank. What to do about non-standard (or non-canonical) language in NLP. *arXiv:1608.07836 [cs]*, August 2016. arXiv: 1608.07836.

Paolo Ramat. *Linguistic typology*, volume 1. Walter de Gruyter, 1987.

Kenji Sagae and Jun'ichi Tsujii. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. In *EMNLP-CoNLL*, volume 2007, pages 1044–1050, 2007.

Edward Sapir. An introduction to the study of speech. *Language*, 1921.

August Wilhelm von Schlegel. *Observations sur la langue et la litterature provencales*. Paris, Librairie grecque-latine-allemande, 1818.

August Schleicher. *Die Sprachen Europas in systematischer Ubersicht: Linguistische Untersuchungen (Bonn, 1850)*, volume 4. John Benjamins Publishing, 1983.

Petr Sgall. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(203-225), 1967.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. A Gold Standard Dependency Corpus for English. In *LREC*, pages 2897–2904, 2014.

Milan Straka, Jan Hajic, and Jana Strakova. UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

David Temperley. Minimization of dependency length in written English. *Cognition*, 105(2):300–333, November 2007.

Lucien Tesniere. *Elements de syntaxe structurale*. Librairie C. Klincksieck, 1959.

Nikolai S. Trubetzkoy. Gedanken ÃŒber das Indogermanenproblem. *Acta linguistica*, 1(1):81–89, 1939.

Reut Tsarfaty, Joakim Nivre, and Evelina Ndersson. Evaluating Dependency Parsing: Robust and Heuristics-free Cross-nnotation Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 385–396, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Erica Tusa, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessit. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, December 2016.

Gertjan Van Noord. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the 10th International Conference on Parsing Technologies*, pages 1–10. Association for Computational Linguistics, 2007.

Viveka Velupillai. *An Introduction to Linguistic Typology*. John Benjamins Publishing Company, Amsterdam ; Philadelphia, August 2012.

G. von der Gabelentz. *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. T.O. Weigel, Leipzig, 1901.

Daniel Zeman. Reusable Tagset Conversion Using Tagset Drivers. In *LREC*, 2008.

Torsten Zesch, Iryna Gurevych, and Max Muhlhauser. Analyzing and accessing Wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications*, pages 197–205, 2007.

Arnold M. Zwicky. Heads. *Journal of linguistics*, 21(01):1–29, 1985.