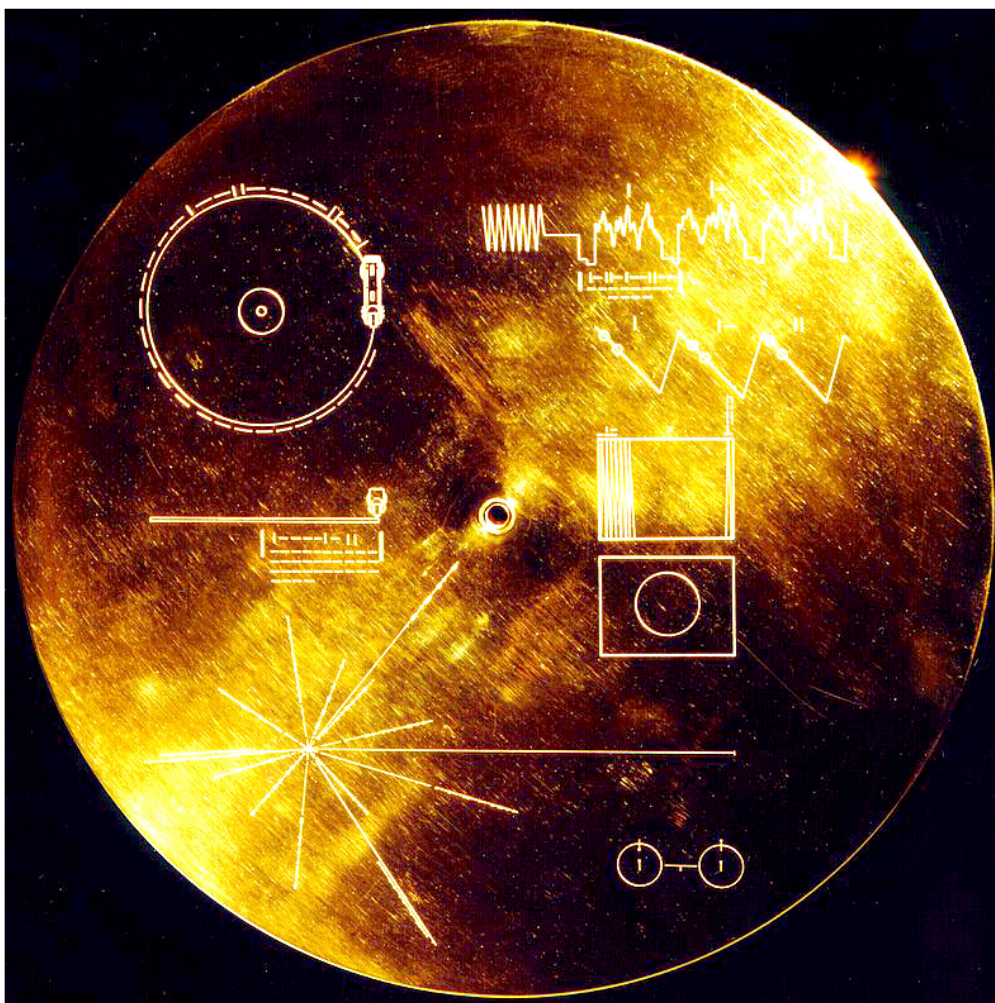




COMPTER PARLER SOIGNER

Tra linguistica e intelligenza artificiale

Atti

Pavia, 15-17 dicembre 2014

a cura di

Edoardo Maria Ponti – Marco Budassi

Pavia University Press
Scientifica

Atti



COMPTON PARLER SOIGNER

Tra linguistica e intelligenza artificiale

Atti

Pavia, Collegio Ghislieri, 15-17 dicembre 2014

a cura di

Edoardo Maria Ponti – Marco Budassi



Pavia University Press

Compter parler soigner : tra linguistica e intelligenza artificiale : atti : Pavia, Collegio Ghislieri, 15-17 dicembre 2014 / a cura di Edoardo Maria Ponti, Marco Budassi. – Pavia : Pavia University Press, 2016. – XIX, 119 p. ; 24 cm.

(Atti)

<http://archivio.paviauniversitypress.it/oa/9788869520389.pdf>

ISBN 9788869520372 (brossura)

ISBN 9788869520389 (e-book PDF)

In testa al front.: Ghislieri.

© 2016 Pavia University Press, Pavia

ISBN: 978-88-6952-038-9

Nella sezione *Scientifica* Pavia University Press pubblica esclusivamente testi scientifici valutati e approvati dal Comitato scientifico-editoriale.

I diritti di traduzione, di memorizzazione elettronica, di riproduzione e di adattamento anche parziale, con qualsiasi mezzo, sono riservati per tutti i paesi.

I curatori sono a disposizione degli aventi diritti con cui non abbiano potuto comunicare, per eventuali omissioni o inesattezze.

In copertina: *cover del disco Golden Record, a bordo della sonda Voyager lanciata nel 1977, contenente saluti in 55 lingue e un messaggio vocale di Jimmy Carter.*

Prima edizione: luglio 2016

Pavia University Press – Edizioni dell'Università degli Studi di Pavia
Via Luino, 12 – 27100 Pavia (PV) – Italia
www.paviauniversitypress.it – unipress@unipv.it

Printed in Italy

Sommario

Introduzione

Edoardo Maria Ponti e Marco Budassi	vii
---	-----

Calcolo di Lambek, logica lineare e linguistica

Claudia Casadio	1
1. Introduzione	1
2. La distinzione funzione–argomento	2
3. Calcolo di Lambek	4
4. Logica Lineare Non-commutativa	7
5. Calcolo di Lambek e logica lineare	11
6. Alcuni esempi linguistici	12
7. Conclusioni	16

Il processamento in tempo reale delle frasi complesse

Cristiano Chesi	21
1. Introduzione	21
2. Una rappresentazione ‘minimalista’ del problema	21
3. Alcune frasi difficili: frasi relative incassate e frasi scisse	24
4. Una rivisitazione <i>processing-friendly</i> delle Grammatiche Minimaliste	32
5. Conclusioni	36

Semantica distribuzionale. Un modello computazionale del significato

Alessandro Lenci	39
1. Introduzione	39
2. Principi e metodi di semantica distribuzionale	40
3. Valutare i modelli semantici distribuzionali	45
4. Le sfide per la semantica distribuzionale	46
5. Conclusioni e prospettive	48

ULISSE: una strategia di adattamento al dominio per l’annotazione sintattica automatica

Felice Dell’Orletta e Giulia Venturi	55
--	----

1.	Introduzione	55
2.	Adattamento al Dominio: definizione del problema	56
3.	Cos'è un dominio?	61
4.	Adattamento al Dominio: metodi	67

Alcuni principi linguistici per costruire risorse lessicali

Elisabetta Jezek	81
1. Introduzione	81
2. La risorsa	81
3. Preferenze di selezione e categorie ontologiche	84
4. <i>Mismatch</i> dei tipi semantici	86
5. Set lessicali instabili	87
6. Conclusioni	89

Alcune note sui Teoremi di Incompletezza di Gödel e la conoscenza delle macchine

Alessandro Aldini, Vincenzo Fano e Pierluigi Graziani	93
1. Gli argomenti gödeliani	93
2. Il punto di vista di Gödel	99
3. Macchine che conoscono se stesse	103
4. Conclusioni	111

<i>Abstracts in English</i>	115
---------------------------------------	------------

ULISSE: una strategia di adattamento al dominio per l'annotazione sintattica automatica

Felice Dell'Orletta e Giulia Venturi – Istituto di Linguistica Computazionale “A. Zampolli”, Consiglio Nazionale delle Ricerche (ILC-CNR) –
felice.dellorletta@ilc.cnr.it giulia.venturi@ilc.cnr.it

1. Introduzione

Fino alla metà degli anni Ottanta gli algoritmi utilizzati per sviluppare sistemi di annotazione linguistica automatica sfruttavano modelli della lingua formati da insiemi complessi di regole scritte a mano e definite a priori a partire dalla conoscenza del sistema linguistico da formalizzare (detti anche ‘modelli simbolici’). Il compito degli algoritmi era quello di identificare le regole da applicare in un determinato contesto e di eseguire la specifica azione. Le principali difficoltà di questa tipologia di algoritmi sono per lo più legate ad aspetti intrinseci del linguaggio naturale, quali l'ambiguità linguistica, la variazione della lingua rispetto al dominio di conoscenza o al genere testuale, le evoluzioni lungo l'asse temporale, ecc.

Basandosi sulla formalizzazione a priori del linguaggio, i modelli simbolici a regole si scontrano pertanto con la difficoltà intrinseca di modellare l'intera conoscenza di una lingua attraverso un insieme finito di regole. L'ostacolo è rappresentato dalla necessità di costruire un modello che sia adattabile alle variazioni linguistiche o all'analisi di *input* ‘non standard’ (es. la lingua dei *social media*, trascrizioni di parlato, ecc.).

Dalla metà degli anni Ottanta grazie al crescente sviluppo di risorse linguistiche, quali ad esempio lessici, *thesauri*, corpora linguisticamente annotati, e al progresso degli studi nel campo dell'Intelligenza Artificiale, iniziano ad essere sviluppati algoritmi basati sull'apprendimento automatico (*Machine Learning*). Tra questi particolare successo hanno avuto gli algoritmi basati su apprendimento supervisionato che risultano particolarmente efficienti ed accurati nella risoluzione di compiti di classificazione. Per poter sfruttare questi algoritmi all'interno dei processi di analisi linguistica, i compiti di annotazione vengono trasformati in compiti di classificazione probabilistica condotta utilizzando un modello statistico estratto da un corpus ‘di addestramento’ nel quale ogni elemento linguistico è associato ad una specifica classe di appartenenza. L'algoritmo è in grado, in fase di addestramento (*training*), di apprendere le varie caratteristiche proprie di un determinato elemento linguistico appartenente ad una classe e di valutare per ognuna di queste caratteristiche il contributo nel determinare la corrispondente classe di appartenenza. Terminata la fase di addestramento, in fase di analisi, l'algoritmo è in grado di estrarre da un nuovo testo preso in esame (non presente nel corpus di addestramento) le varie caratteristiche che contribuiscono a determinare

la classe di appartenenza dell'elemento linguistico. Questi algoritmi sono quindi in grado di estrarre automaticamente dai corpora annotati i modelli linguistici che rappresentano la conoscenza della lingua necessaria per risolvere un determinato compito di analisi linguistica (come ad esempio l'annotazione morfo-sintattica, sintattica, ecc.).

Nonostante questi algoritmi siano più efficienti, accurati e robusti, tuttavia anch'essi soffrono del problema legato alla variazione della lingua rispetto al dominio di conoscenza, al genere testuale, ecc. In questo caso il problema principale è legato alla necessità di adattare il corpus 'di addestramento' alla varietà linguistica in esame. È questo il tema di cui discuteremo in questo contributo finalizzato a definire il problema noto come *Domain Adaptation* e a descrivere in particolare un recente approccio, basato su di un algoritmo chiamato ULISSE (Dell'Orletta et al., 2011), sviluppato dall'ItaliaNLP Lab⁴² dell'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR di Pisa. Questo algoritmo permette di ottenere ottimi risultati nell'adattamento dei sistemi di annotazione linguistica a diverse varietà linguistiche o domini di conoscenza. Nello specifico, affronteremo il caso dell'Adattamento al Dominio (*Domain Adaptation*) di algoritmi di annotazione sintattica automatica dal momento che questa annotazione rappresenta uno dei livelli di analisi linguistica più complessi e più utilizzati in numerosi compiti di Trattamento Automatico del Linguaggio Naturale, come ad esempio estrazione di relazioni, semplificazione automatica, traduzione automatica, ecc. per i quali l'analisi sintattica costituisce il punto di partenza.

2. Adattamento al Dominio: definizione del problema

"Può un corpus annotato [rappresentativo di una certa varietà linguistica] essere utilizzato per l'analisi sintattica di un altro corpus [rappresentativo di un'altra varietà]?"⁴³ La domanda posta nel 2001 da Daniel Gildea ci introduce ad un problema cruciale con cui si devono confrontare i sistemi automatici di analisi sintattica (*syntactic parser* o *parser*) basati su metodi di apprendimento supervisionato: la loro precisione di analisi diminuisce in modo considerevole quando i sistemi vengono applicati su corpora rappresentativi di un dominio diverso rispetto a quello su cui sono stati addestrati. Come riporta Gildea (2001) stesso, il *Collins' parser* (Collins, 1999) addestrato sugli articoli del Wall Street Journal diminuisce la propria accuratezza di analisi di circa 5 punti percentuali quando testato sui testi del Brown Corpus benchè rappresentativi dello stesso genere testuale (giornalistico). O ancora, lo stesso *parser* addestrato sugli articoli della Penn Treebank (Marcus et al., 1993) e testato sul GENIA corpus (Tateisi et al., 2005), un corpus di *abstract* di articoli biomedici, diminuisce di circa 7 punti percentuali (Clegg, Shepherd, 2005). Da qui la necessità di mettere a punto metodi di *Domain Adaptation*, che permettano di contenere la diminuzione dell'accuratezza nell'analisi di testi di nuovi domini.

⁴²Italian Natural Language Processing Laboratory, URL:<www.italianlp.it>

⁴³"Can training data from one corpus be applied to parsing another?" (Gildea, 2001).

La definizione di una strategia per adattare un *parser* statistico ad un nuovo dominio si inserisce nel contesto più ampio delle sfide poste dall'utilizzo di strumenti di Trattamento Automatico del Linguaggio per l'annotazione linguistica automatica di corpora testuali di dominio. A partire dai primi studi condotti negli anni Ottanta, l'attenzione si è in questo senso focalizzata sulla necessità di accordare le varie fasi di elaborazione automatica del testo alle specificità di un determinato linguaggio specialistico.⁴⁴ Dal momento che gli strumenti di analisi automatica del testo sono tipicamente sviluppati per l'elaborazione della lingua usata nei giornali, considerata rappresentativa della lingua comune, l'obiettivo è stato sin d'allora quello di adattare gli strumenti al riconoscimento della struttura linguistica specifica di testi di dominio.

Gli approcci seguiti negli anni Ottanta si basavano sulla possibilità di individuare le peculiarità di un determinato linguaggio specialistico a partire dalle differenze linguistiche esistenti tra lingua standard e di dominio. Tale approccio aveva il suo fondamento nella *Theory of Sublanguages* di Zellig Harris.⁴⁵ Secondo Harris ogni linguaggio specialistico è descrivibile come un 'sottoinsieme' dell'insieme rappresentato dal linguaggio naturale. In analogia con la teoria matematica degli insiemi, esso può dunque essere operativamente denominato *sublanguage*, inteso come "un sottoinsieme del linguaggio comune che si comporta essenzialmente come l'intero linguaggio, essendo però circoscritto a uno specifico dominio" (Grishman, Kittredge, 1986).⁴⁶ Ogni linguaggio specialistico è pertanto studiabile nei termini di restrizione o ampliamento, intersezione o deviazione rispetto alle caratteristiche proprie del linguaggio comune.

Lo studio delle somiglianze e differenze tra un *sublanguage* e la lingua comune aveva un duplice intento, teorico e applicativo. Da un lato, l'analisi dei *sublanguages* ne rendeva interessante lo studio sul piano teorico in quanto "microcosmi dell'intero linguaggio"⁴⁷ in grado di fornire informazioni sul linguaggio naturale stesso (Grishman, Kittredge, 1986). Dall'altro, tale interesse era finalizzato a indagare come le caratteristiche proprie di un *sublanguage* potessero ripercuotersi sull'accuratezza dell'analisi linguistica automatica di testi di dominio.

I domini specialistici oggetto di maggiore interesse erano quelli caratterizzati da un linguaggio altamente specialistico come quello biomedico. È stato infatti questo il dominio di conoscenza al centro delle principali attività di ricerca condotte in quegli anni. Il progetto considerato pioniere in questo ambito, il *Linguistic String Project*,⁴⁸ avviato nel 1965, era espressamente finalizzato a definire una strategia di annotazione sintattica di testi di argomento biomedico, primo passo di un processo completo di trattamento automatico dell'informazione contenuta in testi

⁴⁴Vedi tra gli altri gli studi di Kittredge (1982), Grishman, Nhan et al. (1984), Grishman, Kittredge (1986) e Lehrberger (1986).

⁴⁵Per una descrizione completa della teoria di Harris si rimanda a Harris (1968).

⁴⁶"A subsystem of language that behaves essentially like the whole language, while being limited in reference to a specific subject domain" (Grishman, Kittredge, 1986).

⁴⁷"Microcosms of the whole language" (Grishman, Kittredge, 1986).

⁴⁸Linguistic String Project, URL:<<http://cs.nyu.edu/cs/projects/lsp/index.html>>

di letteratura scientifica (Sager et al., 1987).

Già nell'ambito di questi primi lavori lo studio in particolare delle peculiarità sintattiche rivestiva un interesse specifico. Come ricordato da Bonzi (1990), infatti, "scoprire le regolarità e le principali differenze nell'uso di strutture sintattiche tra diverse discipline e tipologie di testi può aiutare a costruire *parser* più efficienti, può aiutare a scoprire approcci migliori per rintracciare automaticamente i termini che meglio descrivono un documento [...]".⁴⁹ La centralità dell'annotazione sintattica del testo è ricordata più recentemente da Nivre: i risultati dei sistemi di analisi sintattica automatica sono alla base di numerosi compiti di analisi semantica del testo.⁵⁰ Proprio l'importanza dell'analisi sintattica come punto di partenza di analisi più complesse finalizzate all'accesso al contenuto del testo impone che un *parser* per poter essere davvero utile a tal fine debba soddisfare quattro requisiti fondamentali: quelli di *robustness*, *disambiguation*, *accuracy* e *efficiency* (Nivre, 2006).

Oggi sviluppare sistemi che soddisfino tali requisiti è reso ancora più complesso dal fatto che gli strumenti si trovano a dover analizzare non solo testi rappresentativi di diversi domini ma anche 'linguaggi non canonici',⁵¹ come ad esempio trascrizioni del parlato, testi prodotti nei social media (blog, Twitter, Facebook, ecc.), sms, testi di apprendenti e più in generale testi caratterizzati da una sintassi 'non standard'. La difficoltà di sviluppare sistemi di analisi sintattica accurati e robusti ha imposto pertanto la definizione di nuove strategie di adattamento al dominio.

Queste strategie di adattamento variano a seconda del tipo di paradigma che sta alla base del particolare sistema di analisi sintattica automatica utilizzato. Due sono i principali paradigmi: quelli basati su grammatiche (*grammar-driven approach*) e quelli basati su approcci indotti automaticamente dai dati (*data-driven approach*). Secondo la definizione offerta da Nivre (2006), il primo "dipende da un'approssimazione più o meno soddisfacente del linguaggio", approssimazione definita in modo deduttivo dal linguista sulla base delle sue intuizioni; il secondo "dipende da un'inferenza induttiva condotta a partire da un campione più o meno rappresentativo del linguaggio".⁵² Nel primo caso, l'annotazione sintattica del testo è condotta usando grammatiche formali (come ad esempio le 'grammatiche libere da contesto') in grado di definire un linguaggio *L*, dove tale linguaggio è l'insieme delle frasi derivabili nella grammatica data. In questo tipo di approccio analizzare una frase vuol dire derivare tutte (o alcune) le possibili analisi della

⁴⁹"Uncovering the regularities and major differences in the use of syntactic patterns among various disciplines and text types may help to build more efficient parsers for natural language input, to uncover better ways of automatically finding terms which best describe a document [...]" (Bonzi, 1990).

⁵⁰"If we move to applications that require some kind of semantic analysis of individual sentences, the role of parsing becomes more evident" (Nivre, 2006).

⁵¹Per una definizione dei cosiddetti '*Non-Canonical Languages*' si consulti il sito del First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL) tenutosi nel 2012, URL:<<https://sites.google.com/site/sancl2012/>>.

⁵²"Depends on a more or less satisfactory language approximation [...] on inductive inference from a more or less representative language sample" (Nivre, 2006).

frase data la grammatica G . Il compito della disambiguazione (cioè della scelta di un'unica rappresentazione sintattica della frase date tutte le possibili interpretazioni ammesse dalla grammatica) viene condotto attraverso l'utilizzo di regole o restrizioni per permettere al sistema di scegliere una interpretazione piuttosto che un'altra.

Nel secondo caso, l'annotazione sintattica consiste nel generare l'analisi corretta per una frase attraverso un processo di inferenza induttiva a partire da un cosiddetto *training corpus* di riferimento, "un corpus di riferimento nel quale un esperto ha assegnato l'analisi corretta ad ogni segmento rilevante del testo" (Nivre, 2006).⁵³ È centrale pertanto per questa famiglia di strumenti d'analisi la presenza di un corpus annotato manualmente dal quale gli strumenti 'apprendono' ad associare al testo l'informazione linguistica corretta grazie ad un costante processo inferenziale di classificazione probabilistica. Da questo corpus gli strumenti, utilizzando algoritmi di apprendimento automatico, ricavano un modello matematico probabilistico da applicare all'annotazione linguistica di un corpus 'sconosciuto'. In questo caso, il problema della disambiguazione è intrinsecamente affrontato durante la creazione del modello statistico. Il *parser* estrae direttamente dal corpus le statistiche che gli permettono di risolvere il problema della disambiguazione attraverso un processo di assegnamento di un punteggio di probabilità di analisi ad ogni interpretazione possibile della frase.

I due diversi approcci all'analisi sintattica implicano diversi comportamenti dei sistemi in fase di analisi di testi caratterizzati da una sintassi 'non standard'. Entrano in gioco modi diversi di soddisfare due dei requisiti fondamentali che deve avere un *parser* nell'analizzare un testo: quello 1) di essere 'robusto'⁵⁴ nel trattare *input* mal formato o diverso dal linguaggio per l'analisi del quale è stato sviluppato e 2) di essere in grado di 'disambiguare'⁵⁵ tra possibili analisi diverse. Nei *grammar-driven parser*, ogni frase da analizzare deve far parte del linguaggio definito dalla grammatica; questo non accade nei casi di frasi appartenenti a domini diversi rispetto a quelli per i quali la grammatica è stata definita o quando la frase è appunto caratterizzata da una sintassi non standard. Per questo motivo i *parser* basati sulle grammatiche tendono a soffrire di una scarsa robustezza che può essere migliorata attraverso l'introduzione di nuove regole e/o il 'rilassamento' di quelle esistenti. In ogni caso aumentare la copertura e la robustezza di un sistema a regole, oltre ad essere un processo molto complesso, vuol dire aumentare il grado di possibili interpretazioni sintattiche di una frase, rendendo così il pro-

⁵³"I.e., a reference corpus of texts, where each relevant text segment has been assigned its correct analysis by a human expert" (Nivre, 2006).

⁵⁴Il requisito di 'robustezza' (*robustness*) di uno strumento di *parsing* sintattico è così definito: "A system P for parsing texts in language L satisfies the requirement of robustness if and only if, for any text $T = (x_1, \dots, x_n)$ in L , P assigns at least one analysis to every text sentence $x_i \in T$ " (Nivre, 2006, p. 41).

⁵⁵Il requisito di 'disambiguazione' (*disambiguation*) di uno strumento di *parsing* sintattico è così definito: "A system P for parsing texts in language L satisfies the requirement of disambiguation if and only if, for any text $T = (x_1, \dots, x_n)$ in L , P assigns at most one analysis to every text sentence $x_i \in T$ " (Nivre, 2006, p. 42).

blema della disambiguazione un problema molto difficile e computazionalmente oneroso.

Differentemente, 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. Questa caratteristica permette a questi *parser* di analizzare anche frasi non ben fondate, cioè che non rispettano una struttura sintattica definita da una grammatica di riferimento o normativa. Nivre ricorda che, anche a discapito della correttezza d'analisi, l'approccio *data-driven* consente di assegnare sempre comunque almeno un'analisi alla frase in *input*. Questo fa sì che al cambiare di dominio anche testi caratterizzati da un linguaggio 'diverso' da quello del *training* corpus di riferimento siano sempre comunque analizzati. Questo permette di superare il problema della robustezza tipico degli approcci *grammar-driven*, che per riuscire ad analizzare testi di un dominio specialistico (rappresentativi di un linguaggio 'non standard') richiedono un considerevole impegno umano per estendere la grammatica.

Ponendosi in una prospettiva di *Domain Adaptation*, nel caso dei sistemi *grammar-driven* il problema da affrontare è quello della mancanza di 'copertura' della grammatica, la capacità del sistema di tenere conto dei fenomeni sintattici propri del linguaggio che appartiene alla grammatica. Il metodo seguito dai primi studi basati su questo approccio era dunque quello di tracciare "un profilo del sublanguage in grado di descrivere le relative frequenze distribuzionali di frasi e strutture testuali diverse"⁵⁶ rispetto al profilo della lingua comune (Kittredge, 1982). L'attenzione per le caratteristiche di un *sublanguage* era finalizzata alla creazione di una nuova grammatica formale, costruita a partire dagli usi sintattici tipici, diversi da quelli propri della lingua comune. La direzione di ricerca è esposta chiaramente da Grishman, Nhan et al. (1984) che dichiarano l'intenzione di trovare "una procedura euristica in grado di individuare le informazioni di dominio a partire da testi presi come campioni di un sublanguage", una strategia che permettesse di "adattare una grammatica generica alla sintassi di un particolare sublanguage".⁵⁷

Per quanto riguarda gli approcci *data-driven*, data la loro capacità di produrre una rappresentazione sintattica indipendentemente dalla frase da analizzare, il problema dell'adattamento a nuovi domini si trasforma da un problema di robustezza a uno di accuratezza. Per un *parser data-driven* il miglior metodo per ridurre questo problema consiste nella costruzione di una risorsa da usare in fase di addestramento sintatticamente annotata (*treebank*) in modo manuale e composta da testi rappresentativi della particolare varietà testuale o linguistica da analizzare. Questo approccio per l'adattamento, che possiamo definire 'supervisionato', comporta notevoli costi, sia in termini di tempo sia di competenze necessarie per

⁵⁶"a refined sublanguage profile stating the relative frequencies of different sentence and text structures" (Kittredge, 1982).

⁵⁷"a discovery procedure [...] – a procedure which can determine the domain dependent information from sample texts in the sublanguage [...] adapt a broad-coverage grammar to the syntax of a particular sublanguage" (Grishman, Nhan et al., 1984).

l'annotazione manuale. Per questo motivo sono stati sviluppati negli ultimi anni numerosi algoritmi di *Domain Adaptation* definiti 'semi-supervisionati'. L'obiettivo di questi metodi è quello di definire delle strategie di scelta di frasi che, all'interno di un ampio corpus rappresentativo del dominio che si vuole analizzare, contengono caratteristiche tipiche della varietà linguistica del dominio. Dopo essere state selezionate, in base al tipo di algoritmo semi-supervisionato utilizzato, queste frasi vengono corrette manualmente (processo di *Active Learning*) o analizzate automaticamente (processo di *Self-Training*) e vengono aggiunte al *training corpus* originale per introdurre nuova conoscenza nel sistema di analisi.

In questi ultimi anni la maggior parte dei lavori finalizzati allo sviluppo di sistemi per l'analisi sintattica si sono dedicati allo studio di approcci *data-driven*, grazie soprattutto alla loro maggiore robustezza. Saranno infatti questi i sistemi di annotazione a cui faremo riferimento in questo lavoro. In quanto segue, dopo aver mostrato come sia possibile misurare la 'distanza' tra un corpus di addestramento ed il testo da analizzare, quantificandone le differenze rispetto a diversi livelli di descrizione linguistica (Sezione 3.), illustreremo i principali algoritmi di adattamento 'semi-supervisionato' di *parsers data-driven* (Sezione 4.) focalizzandoci in particolare sul recente algoritmo chiamato ULISSE (Dell'Orletta et al., 2011) e mostrandone i risultati ottenuti su domini e lingue diverse. Sebbene i metodi che verranno descritti possano essere applicati, con piccoli accorgimenti, sia a sistemi di analisi che producono rappresentazioni sintattiche a costituenti sia a dipendenze, in questo lavoro ci focalizzeremo, al solo scopo esplicativo, su questi ultimi.

3. Cos'è un dominio?

Come osservato da Gildea (2001), lo studio delle variazioni linguistiche in corpora rappresentativi di varietà linguistiche diverse è di fondamentale importanza nella progettazione e sviluppo di sistemi statistici per l'analisi sintattica. Sottolineando l'importanza degli studi di Douglas Biber (Biber, 1993; Biber et al., 1998) sulle variazioni tra registri, Gildea chiarisce che "le frequenze di occorrenza di varie strutture nel corpus di addestramento si riflettono nel modello probabilistico indotto dal *parser*".⁵⁸ Ponendosi in un'ottica di adattamento al dominio, l'analisi delle caratteristiche linguistiche proprie del corpus di addestramento è dunque di centrale importanza soprattutto allo scopo di metterne in luce le differenze rispetto alle strutture specifiche dei testi che si intende analizzare. Sono infatti proprio queste differenze a causare possibili difficoltà di analisi poiché, non viste in fase di addestramento, non sono presenti nel modello statistico indotto dal *parser*.

Partendo proprio da queste considerazioni, in quanto segue illustriamo una metodologia di analisi finalizzata a descrivere in modo comparativo la distribuzione di alcune caratteristiche lessicali, morfo-sintattiche e sintattiche all'interno di varietà testuali diverse. La metodologia messa a punto ci permette di offrire una definizione operativa di dominio inteso come "una classe di testi definita da

⁵⁸"The frequencies of various structures in training data are reflected in a statistical parser's probability model" (Gildea, 2001).

una distribuzione interna di caratteristiche linguistiche diversa rispetto ad un altro insieme di testi”.

Tale metodologia si ispira alla prospettiva di indagine di Douglas Biber, finalizzata allo studio delle specificità linguistiche proprie di una data varietà linguistica (o registro). Secondo Biber, è un'analisi comparativa tra registro e lingua standard che consente di arrivare a definire cosa significa comune o raro, un'analisi condotta alla ricerca di similitudini e differenze. In quest'ottica, una descrizione completa spesso implica un'analisi composita durante la quale caratteristiche linguistiche acquisite da più livelli di descrizione linguistica svolgono tutte un ruolo importante nella ricostruzione delle dimensioni di variazione sottese tra le varietà (Biber, 1993).

In quanto segue illustreremo e discuteremo un esempio della metodologia di analisi da noi messa a punto. Abbiamo preso in considerazione due domini, quello giuridico e quello biomedico, e due lingue, italiano e inglese. Le caratteristiche rintracciate nei testi rappresentativi di questi due domini sono state messe a confronto con quelle di testi giornalistici considerati rappresentativi della lingua comune (standard). I corpora qui analizzati sono stati precedentemente utilizzati per lo sviluppo di algoritmi di adattamento al dominio di *parser* statistici durante il 'First Shared Task on Dependency Parsing of Legal Texts', tenutosi nell'ambito del 'Workshop on Semantic Processing of Legal Texts' (SPLeT 2012),⁵⁹ e il 'Workshop on Biomedical Natural Language Processing' (BioNLP 2013).⁶⁰ In entrambe le occasioni, i corpora giornalistici sono stati usati per addestrare i *parser* usati, mentre i testi giuridici e biomedici (annotati in modo manuale) sono stati impiegati per verificarne l'accuratezza in uno scenario di adattamento al dominio. Le differenze tra i corpora di lingua standard e quelli 'di dominio' discusse in quanto segue hanno pertanto una duplice lettura: da un lato, ci permettono di definire empiricamente cos'è un 'dominio' rispetto a una varietà di lingua che consideriamo standard; dall'altro, ci permettono di mettere in luce la necessità di sviluppare strategie di adattamento al dominio che colmino la distanza tra il corpus di addestramento e il corpus da analizzare, distanza intesa come diversa distribuzione di eventi linguistici (caratteristiche) che minano la precisione dei *parser*.

Già a partire da caratteristiche di base del testo come la lunghezza della frase (calcolata come il numero medio di *tokens* presenti in una frase) esistono differenze significative tra i testi giuridici e i testi di letteratura biomedica, da una parte, e gli articoli di giornale, dall'altra. Come mostrano infatti le Figure 6. e 7., gli articoli di giornale contengono frasi mediamente più corte di quelle contenute nei testi di dominio. Più in particolare, la Figura 6. mette a confronto corpora in lingua inglese: 1) la porzione di 447.000 *token* della Penn Treebank (PTB) (Marcus et al., 1993) contenente articoli del Wall Street Journal (WSJ) usata in fase di addestramento (*WSJ_gold*), 2) una collezione più ampia (39.285.425 *token*) di WSJ contenente articoli linguisticamente annotati in modo automatico (dunque senza

⁵⁹SPLeT 2012, URL:<<https://sites.google.com/site/splet2012workshop/shared-task>>

⁶⁰BioNLP 2013, URL:<<http://compbio.ucdenver.edu/BioNLP2013/index.shtml>>

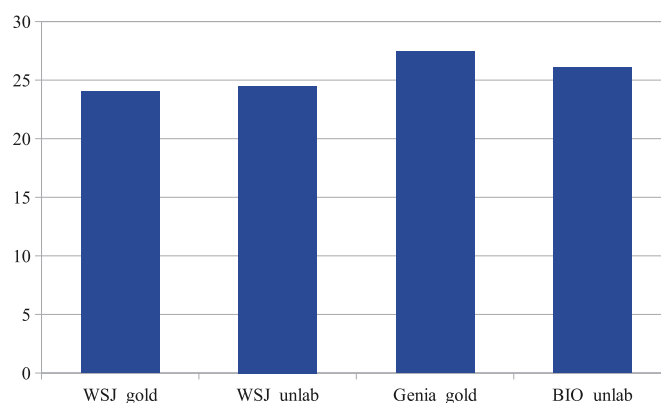
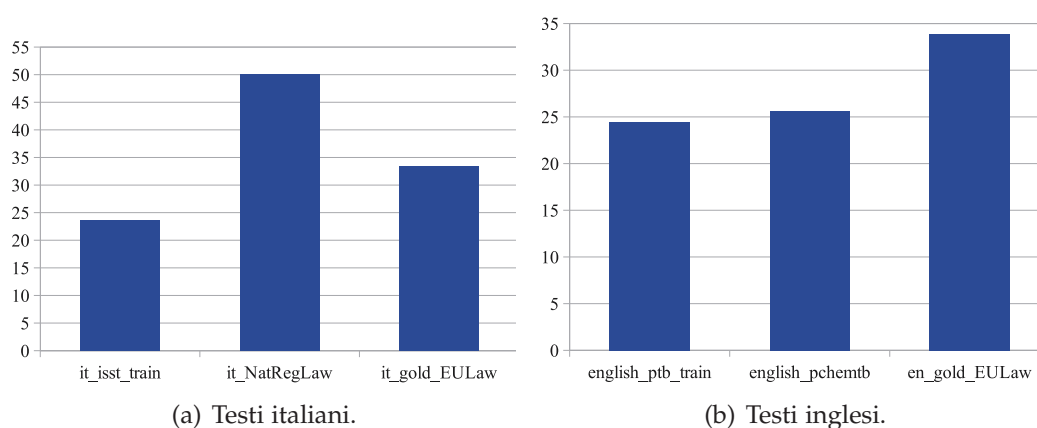


Figura 6. Lunghezza media delle frasi nei testi biomedici e negli articoli di giornale in inglese.



(a) Testi italiani.

(b) Testi inglesi.

Figura 7. Lunghezza media delle frasi nei testi giuridici italiani e inglesi a confronto con gli articoli di giornale nelle due lingue.

revisione manuale) (*WSJ_unlab*), 3) GENIA treebank (Tateisi et al., 2005) composta da *abstract* di articoli biomedici in inglese (circa 493.000 *token*) (*Genia_gold*), 4) un corpus più ampio di articoli di argomento biomedico annotati automaticamente (9.776.890 *token*) (*BIO_unlab*). Come si può notare, le frasi contenute negli articoli di giornale sia annotati in modo manuale (*WSJ_gold*) sia annotati automaticamente (*WSJ_unlab*) sono mediamente più corte di quelle contenute nei testi biomedici (*Genia_gold* e *BIO_unlab*).

Tendenze simili si registrano nell'analisi di testi di dominio giuridico, sia per l'italiano che per l'inglese. La Figura 7.a mostra i confronti tra la lunghezza delle frasi di 1) articoli di giornale contenuti nella porzione della Italian Syntactic-Semantic Treebank (ISST-TANL) (Dell'Orletta, Marchi, Montemagni, Venturi et al., 2013) (71.568 *token*) usata in fase di addestramento dai partecipanti al 'First Shared Task on Dependency Parsing of Legal Texts' (*it_isst_train*), 2) una collezione di te-

sti giuridici emanati dallo stato italiano e dalle regioni (5.194 *token*) (*it_NatRegLaw*) e dalla Commissione Europea (5.662 *token*) (*it_gold_EULaw*), tutti annotati in modo manuale. Anche in questo caso, le frasi contenute negli articoli di giornale sono mediamente più corte di quelle contenute nei testi giuridici e in particolare quelle dei testi giuridici statali e regionali risultano essere le più lunghe.

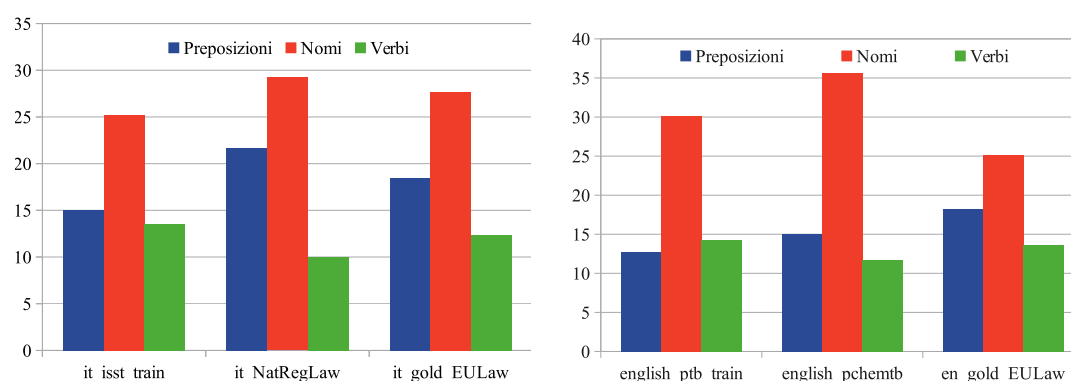
Interessante inoltre come i testi giuridici si caratterizzino per frasi mediamente più lunghe sia dei testi giornalistici sia dei testi biomedici. Questo emerge dall'analisi della Figura 7.b dove sono messi a confronto corpora giornalistici, di testi di argomento biomedico e testi giuridici. Sono in particolare riportati i valori delle lunghezze medie delle frasi contenute 1) nella porzione di articoli giornalistici della PTB usata come corpus di addestramento (446.573 *token*) (*english_ptb_train*), 2) in una collezione di *abstract* biomedici inglesi annotati in modo manuale (5.001 *token*) (*english_pchemtb*), 3) in una collezione di testi giuridici emanati dalla Commissione Europea (5.621 *token*) (*en_gold_EULaw*) anch'essi annotati in modo manuale.

Se ne può dunque concludere che la lunghezza media della frase rappresenta già un primo livello di differenza tra lingua standard e di dominio. È questa una caratteristica che, seppure di base, è strettamente legata ad aspetti di complessità di analisi: se i testi da analizzare contengono frasi più lunghe di quelle contenute nel corpus di addestramento, essi avranno di conseguenza distribuzioni diverse di categorie morfo-sintattiche e caratteristiche sintattiche tipiche di frasi mediamente più lunghe (come ad esempio alberi sintattici più profondi, relazioni di dipendenza sintattica più lunghe, ecc.).

Per quanto riguarda le caratteristiche morfo-sintattiche, i testi del dominio biomedico e giuridico si caratterizzano per una percentuale maggiore di preposizioni e sostantivi e per una minore presenza di verbi (vedi Figura 8.). Come si può notare, tuttavia, vi sono delle differenze significative tra i due domini, ad esempio i testi biomedici si differenziano dai testi giuridici per una percentuale decisamente maggiore di sostantivi. È quello che emerge da un'analisi più dettagliata condotta calcolando la distribuzione delle dieci triple più frequenti di categorie morfo-sintattiche legate da una relazione di dipendenza sintattica⁶¹ nei corpora di testi biomedici (vedi Tabella 4.). Gli *abstract* biomedici (*Genia_gold* e *BIO_unlab* nella Tabella) risultano essere nettamente caratterizzati da una maggiore frequenza di triple di tipo nominale (es. sequenze di due nomi e una preposizione, NN-IN-IN; nome-preposizione-nome, NN-IN-NN; aggettivo-nome-preposizione, JJ-NN-IN) rispetto a quelle di tipo verbale (es. sequenze di determinante-nome-verb, DT-NN-VBZ) rispetto agli articoli di giornale (*WSJ_gold* e *WSJ_unlab* nella Tabella).

Anche a livello delle caratteristiche sintattiche, gli articoli di giornale si differenziano decisamente rispetto ai testi di dominio. Ad esempio sia i testi giuridici che quelli biomedici si caratterizzano per sequenze mediamente più lunghe di complementi preposizionali che modificano a cascata un sostantivo (Figura 9.). Questo tratto risulta specifico in particolare dei testi di dominio giuridico e

⁶¹Ogni tripla è formata da una sequenza di tre categorie morfo-sintattiche legate all'interno di un albero sintattico da relazioni di dipendenza di tipo 'figlio-padre-nonno'.



(a) Categorie morfo-sintattiche in testi italiani. (b) Categorie morfo-sintattiche in testi inglesi.

Figura 8. Distribuzione di categorie morfo-sintattiche nei testi giuridici italiani e inglesi a confronto con gli articoli di giornale nelle due lingue.

WSJ_gold		WSJ_unlab		Genia_gold		BIO_unlab	
Triple	% Freq	Triple	% Freq	Triple	% Freq	Triple	% Freq
DT-NN-IN	2.03	DT-NN-IN	1.72	NN-NN-IN	3.66	DT-NN-IN	2.87
.-VBD-ROOT	1.61	.-VBD-ROOT	1.30	NN-IN-NN	2.93	NN-IN-NN	2.39
NN-IN-NN	1.11	JJ-NN-IN	0.99	DT-NN-IN	2.48	JJ-NN-IN	2.08
JJ-NN-IN	1.10	NN-IN-NN	0.97	JJ-NN-IN	1.96	NN-NN-IN	1.73
.-VBZ-ROOT	1.09	NNP-NNP-IN	0.87	NN-NNS-IN	1.88	IN-NN-IN	1.72
NNP-NNP-IN	0.95	DT-NN-VBD	0.85	JJ-NNS-IN	1.77	JJ-NNS-IN	1.36
DT-NN-VBZ	0.89	NN-VBD-ROOT	0.80	IN-NN-IN	1.65	.-VBD-ROOT	1.33
DT-NN-VBD	0.87	JJ-NNS-IN	0.79	NN-CC-IN	1.64	NNS-IN-NN	1.13
JJ-NNS-IN	0.87	NNP-NNP-VBD	0.78	NNS-IN-NN	1.56	NNP-NN-IN	1.03
IN-NN-IN	0.87	.-VBZ-ROOT	0.75	NN-NN-CC	1.47	NN-IN-VBN	0.93

Tabella 4. Distribuzione percentuale delle dieci triple più frequenti di categorie morfo-sintattiche legate da una relazione di dipendenza sintattica nei testi biomedici e negli articoli di giornale.

soprattutto dei testi giuridici italiani regionali e statali.

Altri due tratti caratterizzanti i domini presi in esame, in particolare i testi giuridici italiani statali e regionali, sono la lunghezza media delle relazioni di dipendenza sintattica (calcolata come la distanza in *token* tra la testa sintattica e il proprio dipendente) (Figura 10.) e la profondità dell'albero sintattico (calcolata come il numero massimo di relazioni di dipendenza consecutive) (Figura 11.). Come messo in evidenza da McDonald e Nivre (McDonald, Nivre, 2007), l'analisi di dipendenze lunghe è una delle principali cause di errore dei sistemi statistici di analisi sintattica. Più in generale, relazioni di dipendenza lunghe e alberi sintattici profondi sono considerati due dei principali elementi di complessità sintattica di una frase (Gibson, 1998; Yngve, 1960; Frazier, 1985). Rispetto a queste osservazioni, i testi dei domini analizzati risultano dunque caratterizzati non solo da eventi linguistici 'distanti' da quelli della lingua comune ma si contraddistinguono anche per una generale maggiore complessità linguistica.

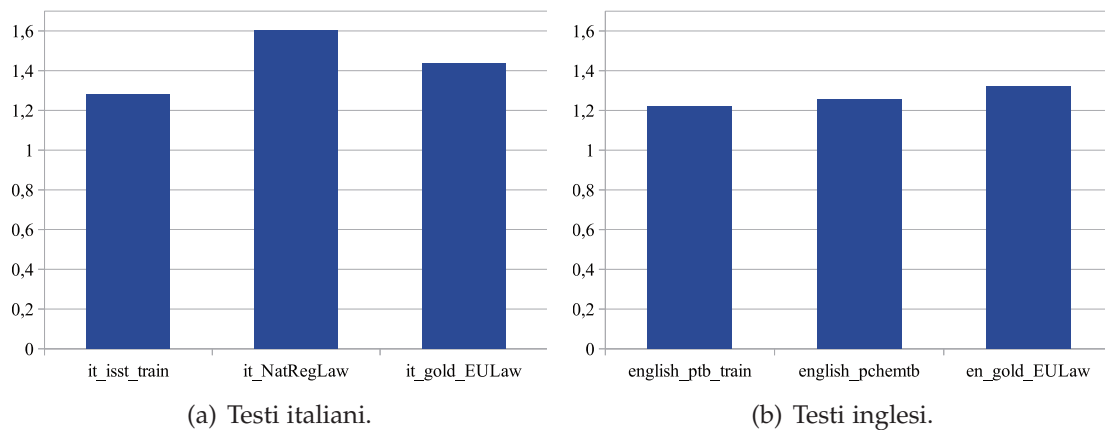


Figura 9. Lunghezza media di complementi preposizionali che modificano un sostantivo nei testi giuridici e biomedici italiani e inglesi a confronto con gli articoli di giornale nelle due lingue.

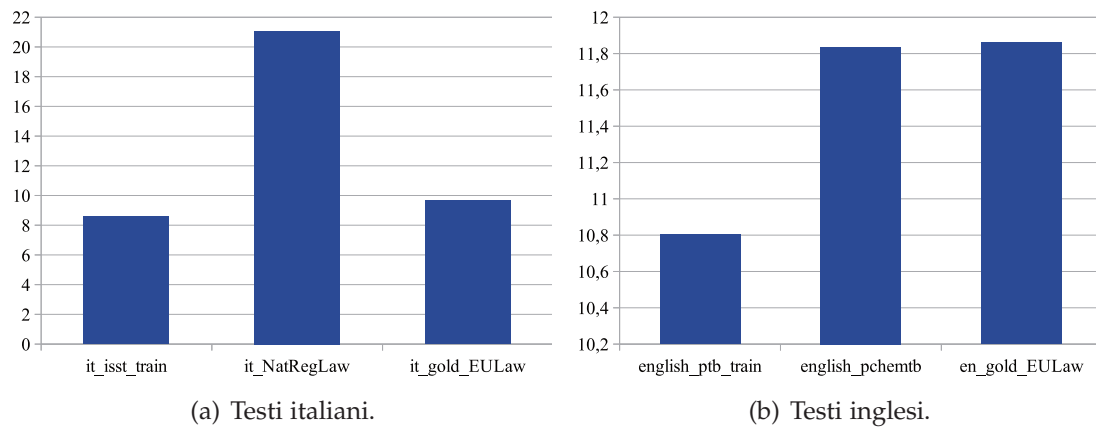


Figura 10. Lunghezza media delle relazioni di dipendenza sintattica nei testi giuridici e biomedici italiani e inglesi a confronto con gli articoli di giornale nelle due lingue.

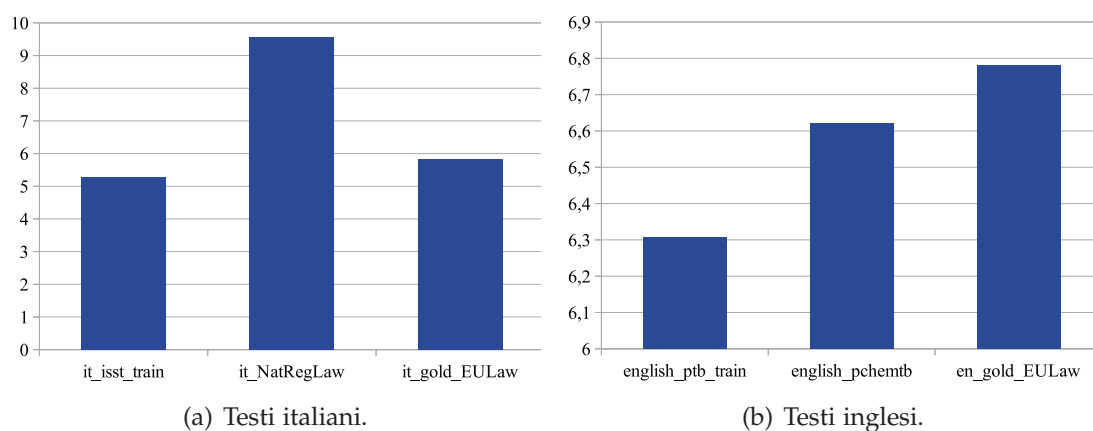


Figura 11. Profondità media dell'albero sintattico nei testi giuridici e biomedici italiani e inglesi a confronto con gli articoli di giornale nelle due lingue.

Veniamo infine alla distribuzione del lessico, un elemento centrale nella definizione di un dominio. Entrambi i domini presi in considerazione contengono una ridotta percentuale dei lemmi presenti nei testi giornalistici usati in fase di addestramento, con alcune differenze significative tra i due domini.⁶² Il corpus di articoli biomedici in inglese e quello di testi giuridici europei in inglese ha una percentuale pari rispettivamente al 60% e all'88% di lemmi presenti nel corpus di testi giornalistici di riferimento così come la collezione di testi giuridici italiani statali e regionali e quella di testi giuridici europei in italiano ha rispettivamente l'81% e l'88% di lemmi del corpus di addestramento. Ne emerge dunque che i due domini si differenziano in modo significativo rispetto al lessico: il dominio giuridico, essendo caratterizzato da un linguaggio che "è un sottinsieme, distinto ma non separato dal linguaggio generale" (Cassese, 1992), si contraddistingue per una maggiore percentuale di lessico giornalistico (cioè comune), caratteristica dei testi giuridici sia italiani sia inglesi. Al contrario, il dominio biomedico è nettamente più distante a livello lessicale dalla lingua giornalistica.

In conclusione, è interessante far notare come anche le differenze nella distribuzione del lessico hanno una diretta ripercussione sull'accuratezza dell'analisi sintattica di testi di dominio. I lemmi sconosciuti, non visti cioè in fase di addestramento, sono quelli per cui il sistema di analisi commetterà il maggior numero di errori, a livello di riconoscimento sia delle categorie morfo-sintattiche sia delle relazioni di dipendenza sintattica nelle quali sono coinvolti.

4. Adattamento al Dominio: metodi

A partire dal 2007 sono state organizzate una serie di conferenze finalizzate a fare il punto sullo stato dell'arte in materia di adattamento al dominio di strumenti di Trattamento Automatico del Linguaggio. Il 'Domain Adptation Track' dello 'Shared Task on Dependency Parsing' organizzato durante la conferenza 'Conference on Computation Natural Language Learning' (CoNLL) è stata la prima occasione in cui si è svolta una competizione tra approcci diversi finalizzati ad adattare un *parser* a domini diversi rispetto a quelli di addestramento, nello specifico testi biomedici, chimici e trascrizioni di parlato infantile (Nivre et al., 2007). In seguito, sono state organizzate numerose altre iniziative tra le quali, nel 2010, il 'Workshop on Domain Adaptation for Natural Language Processing' (DANLP) della 'Conference of the Association of Computational Linguistics' che ha permesso di fare il punto sulle tipologie di approcci di *Domain Adaptation* per diversi compiti di Trattamento Automatico del Linguaggio,⁶³ non solo per sistemi di annotazione sintattica automatica. Più recentemente, nel 2012 in occasione del 'First Workshop on Syntactic Analysis of Non-Canonical Language' (SANCL) e dello 'Shared Task' collegato,⁶⁴ il concetto di dominio è stato allargato alle cosiddette

⁶²In questo caso abbiamo riportato i risultati acquisiti da sezioni dei corpora di dimensioni simili dal momento che la distribuzione del lessico dipende strettamente dalla lunghezza del testo.

⁶³DANLP 2010, URL:<<https://sites.google.com/site/danlp2010/>>

⁶⁴SANCL 2012, URL:<<https://sites.google.com/site/sancl2012/>>

‘varietà non canoniche della lingua’.

Tutte queste iniziative hanno affrontato il tema del *Domain Adaptation* in relazione alla lingua inglese. Poche sono state le occasioni in cui il tema è stato analizzato in rapporto a lingue diverse. Un'eccezione è stato il 'Domain Adaptation Track' (Dell'Orletta, Marchi, Montemagni, Venturi et al., 2013) organizzato in occasione della campagna di valutazione 'Natural Language Processing and Speech Tools for Italian' (EVALITA) nel 2011 e finalizzato a mettere a confronto metodi di adattamento per la lingua italiana. Inoltre, in occasione del 'First Shared Task on Dependency Parsing of Legal Texts' (Dell'Orletta, Marchi, Montemagni, Plank et al., 2012) è stato possibile mettere a confronto metodi di adattamento al dominio per la lingua italiana e inglese.

Come ricordato nella Sezione 2. di questo contributo, la situazione ideale per l'adattamento di un sistema di annotazione sintattica *data-driven* consiste nel poter disporre di una collezione di testi del dominio target annotati in modo manuale. Tuttavia, il principale ostacolo è rappresentato dal fatto che costruire una risorsa sintatticamente annotata in modo manuale è costoso, sia in termini di tempo sia di competenze necessarie. È questo il motivo per cui poche sono le *treebank* oggi esistenti per un dominio diverso da quello giornalistico. I due principali esempi sono: il GENIA corpus (Tateisi et al., 2005), contenente testi biomedici⁶⁵ per un totale di circa 18 mila frasi, e la Google Web Treebank (Petrov, McDonald, 2012) costruita in occasione dello 'Shared Task' del 'First Workshop on Syntactic Analysis of Non-Canonical Language' (SANCL-2012) e contenente diverse tipologie di testi del web (Yahoo answers, e-mail, newsgroup, blog e commenti), per un totale di circa 20 mila frasi.

Data la difficoltà di reperire testi di dominio annotati manualmente, la maggior parte degli studi si è invece concentrata sullo sviluppo di strategie di adattamento semi supervisionato raggruppabili in due classi: quelle che si basano su metodi di *Active Learning* e quelle basate su metodi di *Self-Training*. Entrambe partono da un presupposto comune: il fatto che siano disponibili grandi collezioni di testi di dominio da annotare in modo automatico. L'obiettivo comune è quello di selezionare all'interno di queste grandi collezioni di testi di dominio esempi, cioè frasi 'informative' del dominio target, che possono essere utilizzati in fase di addestramento in aggiunta al corpus di addestramento di partenza.

Diversa è tuttavia la strategia di selezione degli esempi adottata e il modo in cui la loro 'informatività' viene calcolata. I metodi di *Active Learning* si basano su funzioni statistiche in grado di selezionare quegli esempi che, corretti a mano e reinseriti nel corpus di addestramento come nuovo *training*, permettono al sistema di migliorare le proprie accuratèzze. È il caso ad esempio del lavoro di Attardi et al. (2013) che, come criterio di selezione delle frasi da rivedere a mano, hanno utilizzato la probabilità restituita dal classificatore durante la selezione delle azioni del *parser* nella fase di analisi. L'intuizione è stata quella di selezionare come frasi più informative quelle per le quali il *parser* aveva 'più dubbi'

⁶⁵GENIA corpus, URL:<<http://www.nactem.ac.uk/genia/genia-corpus>>

nella scelta dell'interpretazione sintattica. Quelle frasi molto probabilmente erano quelle che contenevano caratteristiche sconosciute rispetto a quelle viste in fase di addestramento.

A differenza dei metodi di adattamento al dominio basati su *Active Learning*, che hanno l'obiettivo di ridurre i tempi e l'impegno di annotazione manuale limitandola ai soli esempi considerati più informativi, i metodi di *Self-Training* non prevedono una fase di revisione manuale. Come nel caso dell'*Active Learning*, questi metodi si differenziano tra loro rispetto alla strategia di selezione delle frasi analizzate automaticamente da inserire nel corpus di addestramento come nuovo *training*. Le frasi selezionate vengono direttamente inserite, senza revisione manuale, come nuovi dati di addestramento in aggiunta a quelli originari. In questo modo il *parser* si 'autoaddestra' su di una collezione nella quale sono uniti esempi annotati a mano (*gold*) ed esempi annotati in modo automatico. In questo caso la difficoltà sta nel definire una funzione che stabilisca sia il livello di 'affidabilità' dell'annotazione automatica sia l'informatività degli esempi da selezionare. I lavori di adattamento al dominio basati su metodi di *Self-Training* (McClosky et al., 2006; Sagae, Tsujii, 2007; McClosky, Charniak, 2008; Sagae, 2010; McClosky et al., 2010 solo per citare alcuni dei più rilevanti) si differenziano pertanto tra loro rispetto alla strategia di selezione automatica dell'affidabilità e dell'informatività degli esempi annotati automaticamente. In quanto segue introdurremo un innovativo metodo di selezione di affidabilità e informatività di frasi annotate in modo automatico (Sezione 4.1.) e ne mostreremo il funzionamento in un sistema di adattamento al dominio basato su *Self-Training* (Sezione 4.2.).

4.1. ULISSE: selezionare alberi sintattici corretti e informativi

Il punto di partenza fondamentale di ogni processo di adattamento al dominio basato su metodi di *Self-Training* è la definizione di una strategia di selezione di alberi sintattici all'interno di un corpus di frasi annotate automaticamente. Per questo motivo si sono diffusi negli ultimi anni diversi approcci finalizzati a stabilire il grado di affidabilità dei risultati dell'annotazione automatica di una frase. Rispetto al metodo utilizzato questi approcci si possono classificare in tre grandi tipi: 1) quelli supervisionati che utilizzano un classificatore esterno per stabilire se le frasi sono state annotate in modo più o meno corretto (Kawahara, Uchimoto, 2008); 2) quelli che combinano i risultati di più *parser* scegliendo le frasi per i quali la maggior parte dei *parser* ha attribuito la stessa analisi (Sagae, Tsujii, 2007); 3) quelli non supervisionati che per la selezione degli alberi sintattici utilizzano solo statistiche acquisite dalle frasi analizzate dal *parser* che si vuole adattare. In questo contributo ci siamo focalizzati su un algoritmo non supervisionato evitando di essere influenzati dalla selezione del corpus di addestramento (metodi supervisionati) e dall'accuratezza dei *parser* utilizzati (metodi che combinano più *parser*). In quanto segue illustreremo un innovativo algoritmo e la sua applicazione in uno scenario di *Self-Training*. Questo algoritmo, chiamato ULISSE, ci ha permesso di migliorare le prestazioni dei *parser* quando applicati su testi di dominio distanti rispetto al loro corpus di addestramento.

ULISSE (*Unsupervised Linguistically-driven Selection of dEpendency parses*) è in grado, data una collezione di frasi automaticamente annotate a livello sintattico a dipendenze, di assegnare ad ogni frase annotata (di qui in avanti riferita come *dependency parse*) un punteggio che quantifica il grado di correttezza dell'analisi automatica. Descriviamo ora il funzionamento dell'algoritmo, mentre per una presentazione più dettagliata rimandiamo al lavoro di Dell'Orletta e colleghi (Dell'Orletta et al., 2011).

La strategia di analisi di ULISSE può essere divisa in due fasi. In una prima fase, ULISSE crea un modello statistico utilizzando un insieme di caratteristiche 'linguisticamente motivate' estratte da un ampio corpus annotato automaticamente. Per la costruzione del modello, ULISSE estrae caratteristiche linguisticamente motivate dai *dependency parse* e da loro sotto-alberi definiti sulla base di criteri 'linguisticamente motivati'. Ad ognuna di esse assegna la probabilità di apparire nel corpus analizzato date alcune 'proprietà' delle frasi all'interno delle quali esse appaiono. Proprietà prese in considerazione sono ad esempio la lunghezza della frase, la distribuzione delle categorie morfo-sintattiche, le parole che le compongono e il loro ordinamento. Vengono quindi acquisite in questa fase le scelte di analisi che il *parser* ha fatto in determinati contesti linguistici. Nella fase successiva, ULISSE assegna un punteggio di accuratezza (di affidabilità dell'analisi) a frasi analizzate automaticamente usando il modello statistico creato nella prima fase. Questo punteggio viene ottenuto calcolando le caratteristiche e i sotto-alberi definiti nella prima fase e presenti nel *dependency parse* preso in considerazione. A queste caratteristiche vengono associati i pesi statistici contenuti nel modello. Una funzione di assegnamento del punteggio combina i pesi per ottenere il risultato finale. L'intuizione su cui si basa ULISSE è che le scelte prese con più frequenza dal *parser* in contesti linguistici simili possono essere considerate più affidabili.

Il punteggio assegnato da ULISSE può essere utilizzato per creare un ordinamento di affidabilità all'interno di un ampio corpus di frasi analizzate. Questo ordinamento può essere considerato il punto di partenza di un processo di *Self-Training*. Oltre alla capacità di creare un ordinamento di affidabilità tra frasi analizzate, altro elemento peculiare di ULISSE è la sua abilità di selezionare le frasi caratterizzanti i fenomeni linguistici tipici del dominio di partenza. Questo è possibile grazie al fatto di fare affidamento su caratteristiche linguisticamente motivate estratte dai *dependency parse*.

La scelta di quali caratteristiche usare per selezionare le frasi corrette e informative è un elemento chiave all'interno dell'intero processo. Tutte le caratteristiche su cui fa affidamento ULISSE sono scelte a partire dalla struttura sintattica a dipendenze e mirano a catturare aspetti dell'annotazione sintattica che caratterizzano in modo particolare il dominio in esame. Questo, unito alla capacità di associare un punteggio di affidabilità all'analisi automatica, rende ULISSE un algoritmo particolarmente adatto per essere usato in un contesto di *Domain Adaptation*.

Possiamo classificare le caratteristiche usate da ULISSE in due tipi: 'globali', calcolate cioè per ogni *dependency parse*, e 'locali', calcolate per ogni singola relazione di dipendenza sintattica. A partire dall'intero albero sintattico vengono calcolate le seguenti caratteristiche che, come abbiamo visto nella Sezione 3.,

rappresentano elementi di variazione tra domini:

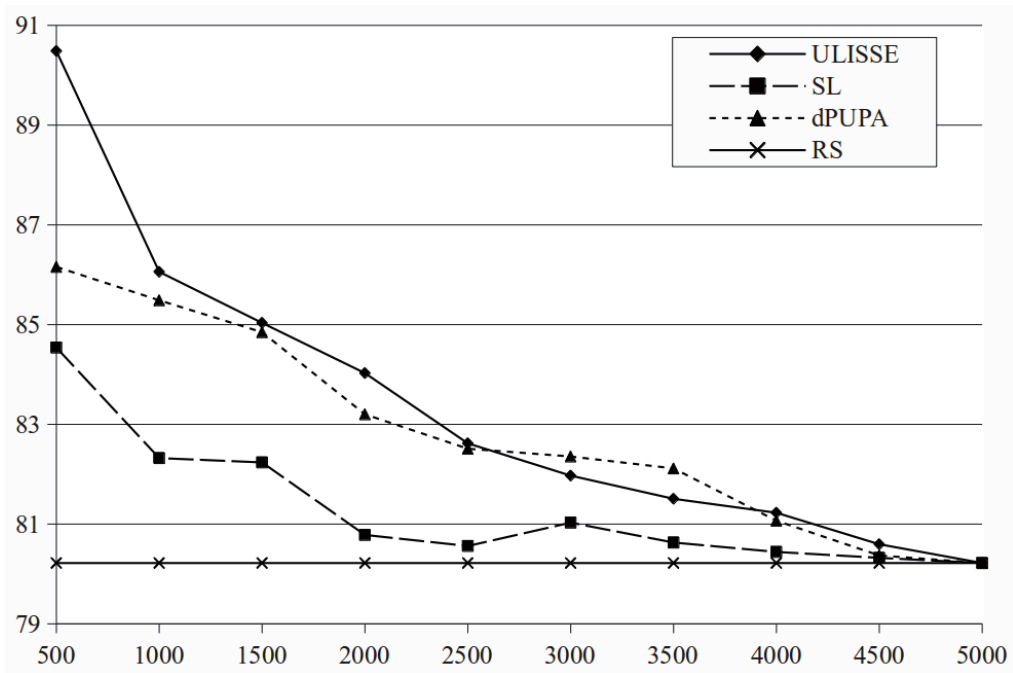
- **profondità media dell'albero sintattico**, calcolata come il numero massimo di relazioni di dipendenza sintattica consecutive;
- **lunghezza media di complementi preposizionali**, calcolata come il numero medio di complementi preposizionali consecutivi che modificano un nome;
- **valenza verbale**, calcolata come il numero di dipendenze sintattiche effettive di un verbo;
- **radici sintattiche verbali**, calcolata come il numero di radici sintattiche di tipo verbale di una frase rispetto al numero totale di radici sintattiche;
- **subordinate contro frasi principali**, calcolata come 1) il rapporto tra frasi subordinate e principali e 2) posizione delle subordinate rispetto alla principale all'interno di una frase;
- **lunghezza delle relazioni di dipendenza sintattica**, calcolata come la distanza tra la testa e il dipendente (in *token*).

Queste caratteristiche, oltre a delineare i tratti linguistici peculiari del dominio analizzato, sono fortemente connesse con la misura della complessità sintattica di una frase. Ad esempio, la profondità dell'albero sintattico di una frase può essere considerata un indicatore di crescente complessità sintattica (Yngve, 1960; Frazier, 1985; Gibson, 1998), così come la posizione della subordinata rispetto alla principale: frasi che contengono subordinate in posizione post-verbale sono più semplici di frasi in cui la subordinata precede la frase principale (Miller, Weinert, 1998). McDonald, Nivre (2007) mostrano che i *parsers* statistici hanno una diminuzione di accuratezza quando si trovano a dover analizzare frasi che contengono relazioni di dipendenza molto lunghe.

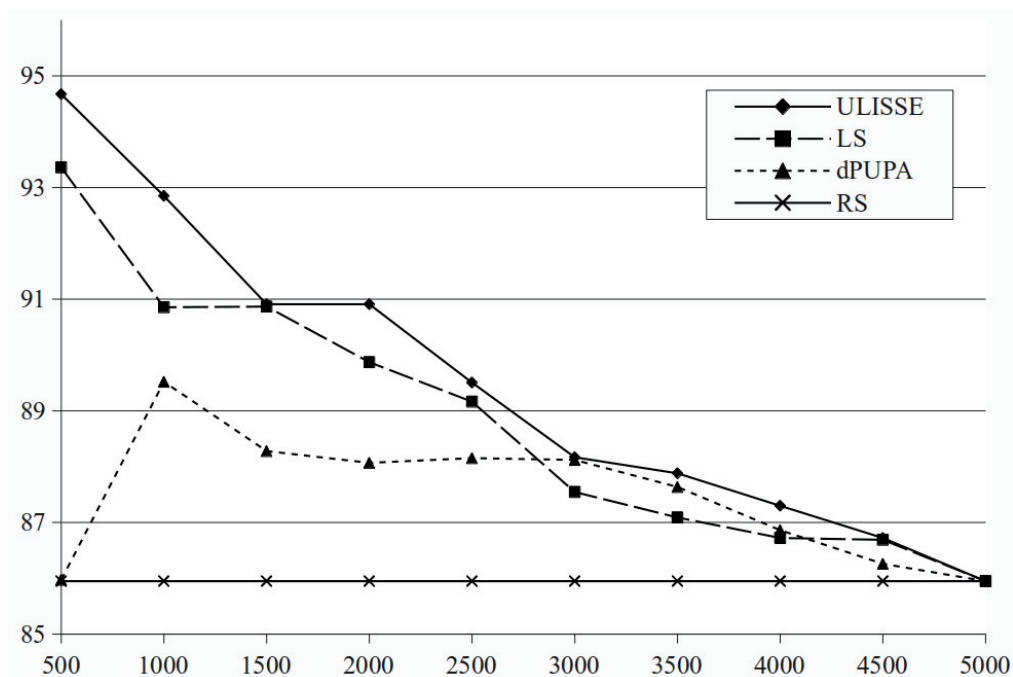
La caratteristica 'locale' di ULISSE associa ad ogni relazione di dipendenza un punteggio di plausibilità. Tale punteggio viene calcolato tenendo in considerazione per ogni relazione: la categoria morfo-sintattica (*part-of-speech*) del dipendente, della testa sintattica, e del padre della testa sintattica, e la relazione sintattica che li lega. Ad ogni *dependency parse* viene associato il peso della relazione sintattica meno plausibile.

Per dimostrare l'efficacia di ULISSE nel predire l'accuratezza dell'analisi sintattica fatta da un *parser*, riportiamo gli esperimenti descritti in Dell'Orletta et al. (2011) dove ULISSE è stato testato su corpora di testi italiani e inglesi annotati automaticamente utilizzando il *parser* DeSR (Attardi, 2006). In tutti gli esperimenti ULISSE si è dimostrato più efficace rispetto agli altri algoritmi di selezione presi come riferimento (*baseline*).⁶⁶

⁶⁶In Dell'Orletta et al. (2011) è possibile trovare altri esperimenti realizzati con ULISSE su diversi domini e utilizzando diversi *parser* che dimostrano l'indipendenza della sua efficacia rispetto alle lingue, ai domini e alle diverse strategie di analisi sintattica utilizzate.



(a) Testi italiani.



(b) Testi inglesi.

Figura 12. Accuratezza di analisi (LAS) di gruppi di frasi annotate in modo automatico e ordinate da ULISSE e dalle *baselines*.

La Figura 12. mostra i risultati degli esperimenti. In particolare, la Figura 12.a riporta i risultati degli esperimenti condotti su un corpus di testi italiani e la Figura 12.b quelli condotti su un corpus di testi inglesi. Per la lingua italiana il *parser* DeSR è stato addestrato sulla ISST-TANL, una *treebank* di articoli giornalistici e periodici di 3.109 frasi (71.285 *token*) usata nella campagna di valutazione di strumenti di Trattamento Automatico del Linguaggio per l'italiano Evalita'09⁶⁷ (Bosco et al., 2009). Come ampio corpus di riferimento da cui ULISSE ha estratto il modello statistico è stata usata una porzione del corpus di articoli giornalistici CLIC-ILC (Marinelli et al., 2003) pari a 1.104.237 frasi (22.830.739 *token*). Il *test set* usato per valutare l'accuratezza di DeSR è quello distribuito in occasione di Evalita'09 composto da 260 frasi (5.011 *token*). Per gli esperimenti condotti sull'inglese, DeSR è stato addestrato sulle sezioni 2–11 della parte della Penn Treebank (Marcus et al., 1993) usata in occasione del 'CoNLL 2007 Shared Task on Dependency Parsing' (Nivre et al., 2007) che comprende 447.000 *token* e circa 18.600 frasi. Come corpus di riferimento è stato usato l'ampio corpus distribuito in occasione del 'CoNLL 2007 Shared Task on Dependency Parsing', composto da 1.625.606 frasi (39.285.425 *token*), e come *test set* la sezione 23 della parte della Penn Treebank pari a 214 frasi (5.003 *token*).

In ogni figura, l'asse delle ordinate riporta l'accuratezza di analisi del *parser* espressa in *Labelled Attachment Score* (LAS), che misura la percentuale di *token* ai quali il *parser* ha assegnato correttamente sia la testa sintattica sia il tipo di relazione di dipendenza. Sull'asse delle ascisse sono riportate le porzioni crescenti dei primi k *token* della lista di *dependency parse* ordinati da ULISSE e le *baseline*. Il valore di k progredisce per intervalli di 500 *token* e va dai primi 500 *token* all'intera porzione del *test set* (dunque $k=500$, $k=1000$, $k=1500$, ecc.). La linea etichettata con *RS* rappresenta l'accuratezza di DeSR sull'intero *test set*, quella etichettata come *ULISSE* rappresenta l'ordinamento prodotto da ULISSE, la linea *SL* rappresenta l'accuratezza ottenuta utilizzando la lunghezza della frase come criterio per ordinare i *dependency parse*. In base a questa *baseline* le frasi annotate sono state ordinate dalle più corte alle più lunghe, assumendo che le più corte fossero le più corrette perché, come dimostrato da McDonald, Nivre (2007), frasi più lunghe sono più difficili da analizzare. La lunghezza della frase come criterio di selezione dei *dependency parse* più affidabili costituisce pertanto una *baseline* piuttosto efficace. Le performance di ULISSE sono inoltre state confrontate con *dPUPA*, che rappresenta la versione adattata per l'analisi sintattica a dipendenze di un algoritmo di ordinamento sviluppato per l'analisi sintattica a costituenti (Reichart, Rappoport, 2009). *PUPA* rappresenta lo stato dell'arte degli approcci di selezione non supervisionati.

Esaminando la Figura 12. possiamo notare come, nonostante la bontà delle *baseline* scelte, per entrambe le lingue ULISSE risulti il criterio di ordinamento migliore. Sebbene la selezione dei *dependency parse* guidata dalla lunghezza della frase risulti un buon criterio di ordinamento, ULISSE è stato in grado di selezionare frasi più lunghe che hanno un livello di accuratezza maggiore rispetto a frasi

⁶⁷Evalita'09, URL:<<http://evalita.fbk.eu/index.html>>

molto brevi. Ad esempio, se osserviamo l'esperimento sull'italiano, mentre le frasi contenute nella prima porzione di 500 *token* ordinate con *SL* hanno una lunghezza media di 7,88 *token* (e una accuratezza di 84,54% di LAS), ULISSE è stato in grado di ordinare nella stessa porzione dei primi 500 *token* frasi con una lunghezza media di 9,5 *token* (e un'accuratezza di 90,48%). Nella stessa porzione di testo *PUPA* ha selezionato frasi con una lunghezza media di 12 *token* ma con un'accuratezza minore di quella ottenuta da ULISSE (86,16%), dimostrandosi tuttavia un ottimo criterio di selezione.

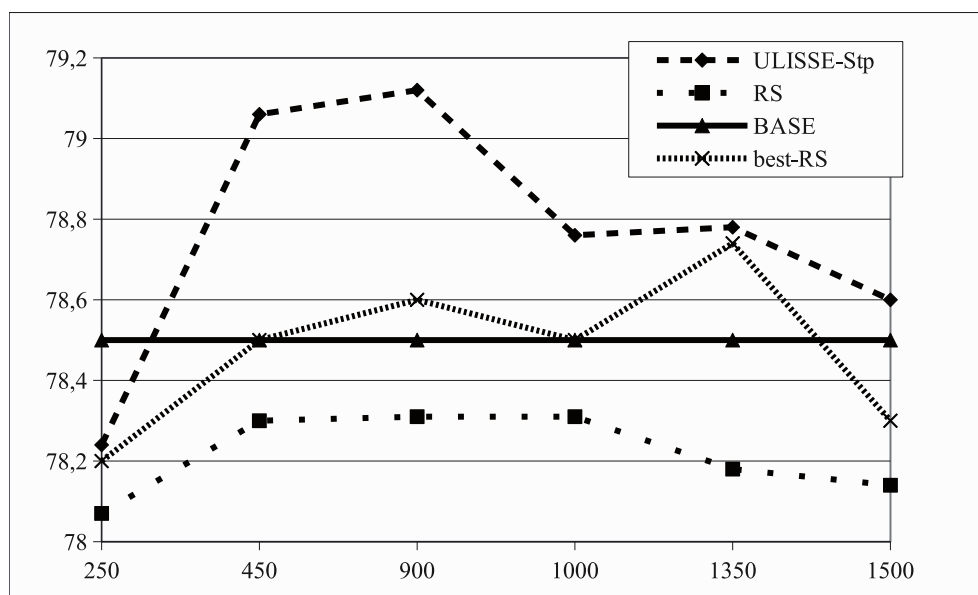
Abbiamo quindi visto come ULISSE sia in grado di selezionare e ordinare per affidabilità decrescente di analisi le frasi all'interno di un ampio corpus facendo unicamente riferimento alle caratteristiche specifiche del corpus stesso. Gli esperimenti mostrati sono stati condotti in uno scenario che possiamo definire *in-domain*, dal momento che il dominio dei testi usati in fase di addestramento del *parser* è lo stesso di quello dell'ampio corpus di documenti su cui è stato testato ULISSE. Sia per l'italiano sia per l'inglese gli esperimenti sono stati condotti su testi giornalistici. I buoni risultati raggiunti ci hanno suggerito come l'algoritmo di selezione potesse essere utilizzato con successo anche nel caso in cui il corpus di riferimento usato da ULISSE fosse diverso da quello su cui il *parser* era stato originariamente addestrato, scenario che possiamo definire *out-domain*. Questo, unito alla sua capacità di estrarre frasi rappresentative del dominio analizzato (grazie all'uso di caratteristiche linguisticamente motivate), fanno di ULISSE un algoritmo particolarmente adatto per essere usato in un contesto di *Self-Training per Domain Adaptation*.

4.2. ULISSE per l'Adattamento al Dominio

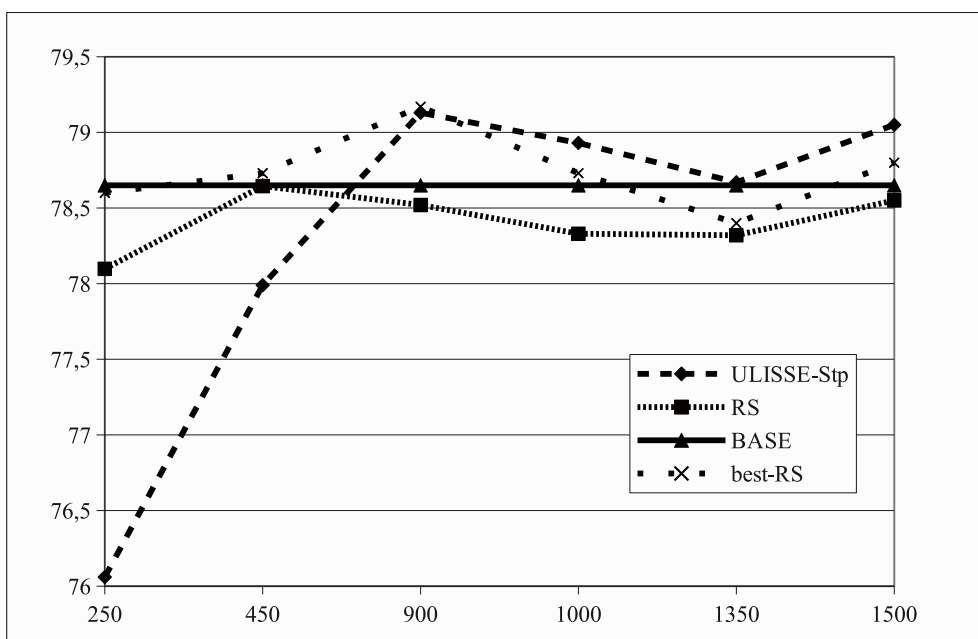
In quanto segue illustreremo alcuni esperimenti condotti usando ULISSE in uno scenario *out-domain*. Gli esperimenti, originariamente condotti da Dell'Orletta et al. (2013) hanno dimostrato come l'algoritmo sia in grado di selezionare e ordinare frasi annotate automaticamente sufficientemente affidabili e informative per essere direttamente inserite nel corpus di addestramento originario affinché il *parser* si 'autoaddestri' adattandosi ad un nuovo dominio. In questo paragrafo ci soffermeremo sugli esperimenti condotti su testi scritti in lingua inglese.

Le Figure 13. e 14. riportano i risultati degli esperimenti di adattamento al dominio condotti con l'obiettivo di adattare il *parser* DeSR all'analisi di *abstract* di argomento biomedico e chimico in inglese. Il *parser* è originariamente addestrato sulle sezioni 2–11 della parte della Penn Treebank (Marcus et al., 1993). Sono stati usati come corpora di riferimento, da cui ULISSE ha potuto estrarre il modello statistico, le collezioni distribuite in occasione del 'CoNLL 2007 Shared Task on Domain Adaptation' (Nivre et al., 2007). Si tratta di una collezione di *abstract* di argomento chimico (di qui in avanti riferiti come *CHEM*) composta da 396.128 frasi (10.482.247 *token*) e di una collezione di *abstract* di argomento biomedico (di qui in avanti riferiti come *BIO*) composta da 375.421 frasi (9.776.890 *token*). Gli esperimenti condotti su *CHEM* sono stati testati su di una porzione di 195 frasi (5.001 *token*) e quelli condotti su *BIO* su una collezione di 200 frasi (5.017 *token*). Le performance di DeSR 'autoaddestrato' usando i *dependency parse* selezionati da

ULISSE sono state messe a confronto con i risultati ottenuti 1) da DeSR addestrato solo sulla Penn Treebank (*BASE* nelle figure), 2) usando in fase di addestramento frasi annotate automaticamente e selezionate in modo casuale nell'ampio corpus di dominio (*RS*) sulla base della media di un processo di *10-fold cross validation*, 3) usando il miglior risultato ottenuto nel processo di *10-fold cross validation* (*best-RS*).



(a) *Abstract* di argomento chimico.



(b) *Abstract* di argomento biomedico.

Figura 13. Accuratezza di analisi (LAS) in uno scenario di *Self-Training* usando le frasi ordinate da ULISSE e dalle *baselines*.

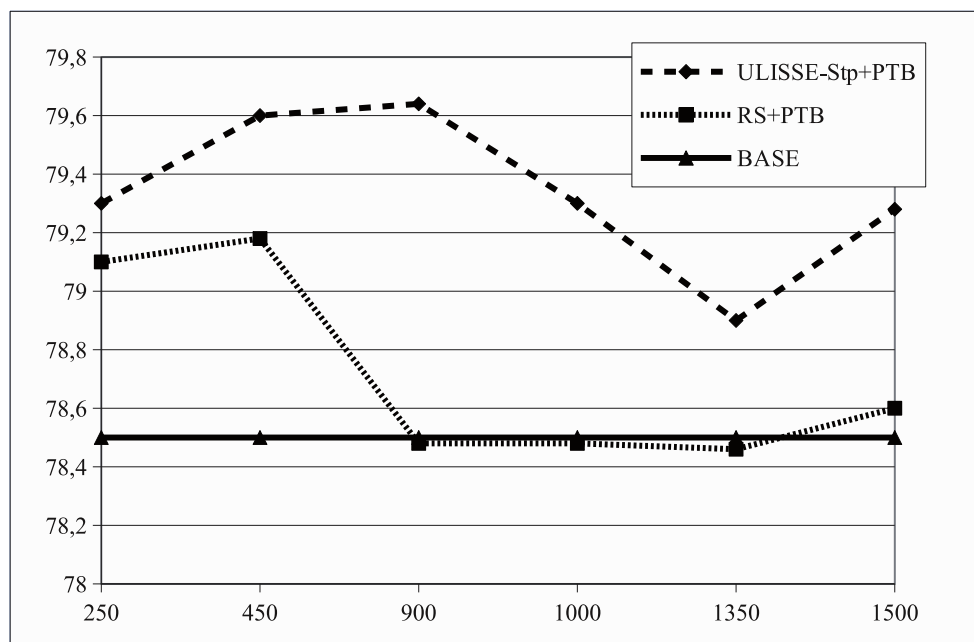
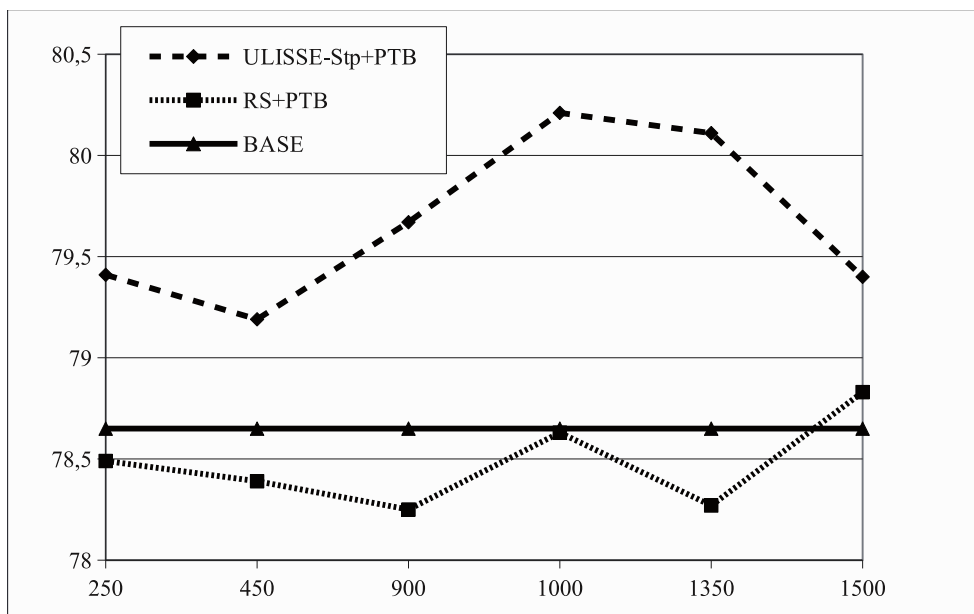
(a) *Abstract di argomento chimico.*(b) *Abstract di argomento biomedico.*

Figura 14. Accuratezza di analisi (LAS) in scenari di *Self-Training* usando le frasi ordinate da ULISSE e dalle *baselines* in aggiunta alle frasi della Penn Treebank.

Sono stati condotti due tipi di esperimenti: un primo nel quale l'accuratezza di DeSR è stata testata su *CHEM* e *BIO* usando in fase di addestramento solo le frasi selezionate da ULISSE e dalle tre *baselines* (si tratta dunque di un corpus

di addestramento costruito in modo completamente automatico); un secondo nel quale a queste frasi sono state aggiunte quelle del corpus di addestramento di partenza di DeSR (in questo caso il corpus di addestramento è composto da una parte annotata in modo manuale e da una in modo automatico). In ogni figura, l'asse delle ordinate riporta l'accuratezza di analisi del *parser* espressa in *Labelled Attachment Score* (LAS), mentre sull'asse delle ascisse sono riportate le dimensioni in *token* delle frasi selezionate da ULISSE e dalle *baseline* (250 migliaia di *token*, 450, 900, ecc.) progressivamente aggiunte in fase di addestramento.

Nel primo tipo di esperimenti (Figure 13.a e 13.b), ULISSE è risultato il miglior algoritmo di *Self-Training*, sebbene nell'esperimento di adattamento al dominio degli *abstract* di argomento biomedico solo aggiungendo 900 mila *token* selezionati da ULISSE l'accuratezza di DeSR migliora rispetto ai risultati raggiunti aggiungendo frasi selezionate dalle *baselines*. In generale, si può comunque notare che 900 mila frasi di addestramento è la quantità massima dopo la quale l'accuratezza del *parser* inizia a decrescere. Ciò significa che dopo un certo numero di nuove frasi ULISSE non riesce più a selezionare dall'ampio corpus annotato automaticamente di partenza *dependency parse* che abbiano un buon livello di affidabilità e siano sufficientemente informative. È interessante notare come DeSR addestrato sui soli *dependency parse* selezionati da ULISSE abbia un'accuratezza maggiore di quella ottenuta usando i testi giornalistici (annotati in modo manuale) della Penn Treebank.

Dai risultati ottenuti nel secondo tipo di esperimenti (Figure 14.a e 14.b), emerge che le frasi selezionate da ULISSE, unite a quelle del corpus di addestramento originale, sono quelle che permettono a DeSR di avere le accuratezze maggiori. Questo risultato dimostra come l'algoritmo di selezione sia non solo in grado di scegliere i *dependency parse* più affidabili ma anche quelli che contengono le specificità del dominio. Sono queste le frasi che, essendo rappresentative degli eventi linguistici tipici del dominio da analizzare, aggiungono nuova informazione al corpus di addestramento originario permettendo a DeSR di adattarsi.

Bibliografia

- Attardi G., M. Simi, A. Zanelli (2013). "Domain Adaptation by Active Learning". *Evaluation of natural language and speech tool for Italian*. B. Magnini, F. Cutugno, M. Falcone, E. Pianta (cur.) Heidelberg: Springer-Verlag, pp. 77–85.
- Attardi G. (2006). "Experiments with a multilanguage non-projective dependency parser". *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, pp. 166–170.
- Biber D. (1993). "Using register-diversified corpora for general language studies". *Computational linguistics journal*. Vol. 19(2), pp. 219–241.
- Biber D., S. Conrad, R. Reppen (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.
- Bonzi S. (1990). "Syntactic patterns in scientific sublanguages: a study of four disciplines". *Journal of the American Society for Information Science*, 41 (2), pp. 121–131.

- Bosco C., S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, A. Lenci (2009). "Parsing Task: comparing dependency parsers and treebanks." *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, [senza numerazione].
- Cassese S. (1992). "Introduzione allo studio della normazione". *Rivista trimestrale di diritto pubblico*, 2, pp. 307–330.
- Clegg A. B., A. J. Shepherd (2005). "Evaluating and integrating treebank parsers on a biomedical corpus". *Proceedings of the Workshop on Software*, pp. 14–33.
- Collins M. (1999). "Head-driven statistical models for natural language parsing". Tesi di dott. University of Pennsylvania.
- Dell'Orletta F., S. Marchi, S. Montemagni, B. Plank, G. Venturi (2012). "The SPLeT-2012 shared task on dependency parsing of legal texts". *Proceedings of the Fourth Workshop on Semantic Processing of Legal Texts (SPLeT 2012)*, pp. 42–51.
- Dell'Orletta F., S. Marchi, S. Montemagni, G. Venturi, T. Agnoloni, E. Francesconi (2013). "Domain adaptation for dependency parsing at Evalita 2011". *Evaluation of natural language and speech tool for Italian*. B. Magnini, F. Cutugno, M. Falcone, E. Pianta (cur.) Heidelberg: Springer-Verlag, pp. 58–69.
- Dell'Orletta F., G. Venturi, S. Montemagni (2011). "ULISSE: an unsupervised algorithm for detecting reliable dependency parses". *Proceedings of CoNLL 2011*, pp. 115–124.
- Dell'Orletta F., G. Venturi, S. Montemagni (2013). "Unsupervised linguistically-driven reliable dependency parses detection and self-training for adaptation to the biomedical domain". *Proceedings of the 12th workshop on Biomedical Natural Language Processing (BioNLP-2013)*, pp. 45–53.
- Frazier L. (1985). "Syntactic complexity", in L. Karttunen D.R. Dowty, A.M. Zwicky (cur.), *Natural language parsing*. Cambridge: Cambridge University Press.
- Gibson E. (1998). "Linguistic complexity: locality of syntactic dependencies". *Cognition*. Vol. 68, pp. 1–76.
- Gildea D. (2001). "Corpus variation and parser performance". *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, pp. 167–202.
- R. Grishman, R. Kittredge (cur.), cur. (1986). *Analyzing language in restricted domains: sublanguage description and processing*. Hillsdale: Lawrence Erlbaum.
- Grishman R., N. T. Nhan, E. Marsh, L. Hirschman (1984). "Automated determination of sublanguage syntactic usage". *Proceedings of the 10th International Conference on Computational Linguistics*, pp. 96–100.
- Harris Z. (1968). *Mathematical Structures of Language*. Hoboken: Wiley.
- Kawahara D., K. Uchimoto (2008). "Learning reliability of parses for domain adaptation of dependency parsing". *Proceedings of IJCNLP 2008*, pp. 709–714.
- Kittredge R. (1982). "Variation and homogeneity of sublanguages", in R. Kittredge, J. Lehrberger (cur.), *Sublanguage: studies of language in restricted semantic domains*. Berlin: De Gruyter, pp. 107–137.
- Lehrberger J. (1986). "Sublanguage analysis". *Analyzing language in restricted domains: sublanguage description and processing*. R. Grishman, R. Kittredge (cur.) New York: Lawrence Erlbaum, pp. 19–38.

- Marcus M. P., M. A. Marcinkiewicz, B. Santorini (1993). "Building a large annotated corpus of English: The Penn Treebank". *Computational linguistics*, 19.2, pp. 313–330.
- Marinelli R., L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari, A. Zampolli (2003). "The Italian PAROLE corpus: an overview", in A. Zampolli (cur.), *Linguistica computazionale*. Pisa: Giardini, pp. 401–421.
- McClosky D., E. Charniak (2008). "Self-training for biomedical parsing". *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pp. 101–104.
- McClosky D., E. Charniak, M. Johnson (2010). "Automatic domain adaptation for parsing". *Proceedings of the HLT-NAACL'2010*, pp. 28–36.
- McClosky D., E. Charniak, M. Johnson (2006). "Reranking and self-training for parser adaptation". *Proceedings of ACL 2006*, pp. 337–344.
- McDonald R., J. Nivre (2007). "Characterizing the errors of data-driven dependency parsing models". *Proceedings of the the EMNLP-CoNLL*, pp. 122–131.
- Jim Miller, Regina Weinert (cur.), cur. (1998). *Spontaneous spoken language. Syntax and discourse*. Oxford: Clarendon Press.
- Nivre J. (2006). *Inductive dependency parsing*. Berlin: Springer.
- Nivre J., J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel, D. Yuret (2007). "The CoNLL 2007 Shared Task on Dependency Parsing". *Proceedings of the the EMNLP-CoNLL*, pp. 915–932.
- Petrov S., R. McDonald (2012). "Overview of the 2012 shared task on parsing the web". *Proceedings of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL 2012)*, pp. 1–8.
- Reichart R., A. Rappoport (2009). "Automatic selection of high quality parses created by a fully unsupervised parser". *Proceedings of CoNLL 2009*, pp. 156–164.
- Sagae K. (2010). "Self-training without reranking for parser domain adaptation and its impact on semantic role labeling". *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP 2010)*, pp. 37–44.
- Sagae K., J. Tsujii (2007). "Dependency parsing and domain adaptation with LR models and parser ensemble". *Proceedings of EMNLP-CoNLL 2007*, pp. 1044–1050.
- N. Sager, C. Friedman, M. Lyman (cur.), cur. (1987). *Medial language processing*. Reading: Addison-Wesley Publishing Company.
- Tateisi Y., A. Yakushiji, T. Ohta, J. Tsujii (2005). "Syntax annotation for the GENIA corpus". *Companion volume to Proceedings of IJCNLP'05*, pp. 222–227.
- Yngve V. H. (1960). "A model and an hypothesis for language structure". *Proceedings of the American Philosophical Society*, 104.5, pp. 444–466.