UNIVERSITÀ DI PISA

FACOLTÀ DI LETTERE E FILOSOFIA

Corso di Laurea in Informatica Umanistica

TESI DI LAUREA

Adattamento del parsing sintattico a testi di dominio giuridico

Relatore

Prof. Alessandro Lenci

Candidato

Tommaso Petrolito

Anno Accademico 2011-2012

Indice

Introduzione e ringraziamenti

L'idea di analizzare questo aspetto dell'annotazione linguistica automatica, parte dal lavoro che ho svolto all'Istituto di Linguistica Computazionale del CNR di Pisa per il tirocinio previsto dal piano di studi in Informatica Umanistica.
Il tirocinio "Annotazione di frasi estratte da corpora giuridici con analisi sintattiche e frame semantici" mi ha fornito l'opportunità di addentrarmi in questo campo di ricerca e di analizzare con interesse le criticità e le problematicità del caso specifico di annotazione di corpora di dominio giuridico.
Analizzando e correggendo manualmente l'output elaborato dal parser ho potuto rendermi conto direttamente dei problemi di annotazione automatica (sia morfosintattica che sintattica) specifici del dominio giuridico già evidenziati da Venturi (2011).
L'aggiunta di piccoli set di frasi annotate manualmente al corpus di addestramento ha progressivamente migliorato i risultati dell'annotazione.
In particolare è risultato evidente che ad un iniziale elevato numero di errori poco significativi, man mano è seguito un numero progressivamente minore di errori più significativi.
Una curiosità è nata spontaneamente al termine del lavoro: l'aggiunta di questo piccolo set di frasi annotate manualmente ha migliorato e migliora in generale i risultati dell'annotazione su corpora di dominio giuridico, ma è da accertare se migliori, peggiori, o lasci invariati i risultati dell'annotazione su corpora della lingua comune, quali ad esempio corpora di linguaggio giornalistico.

Questo lavoro è il coronamento dei mie 3 anni di studi in Informatica Umanistica presso l'Università di Pisa.
Ringrazio, innanzi tutto, Alessandro Lenci per avermi fatto conoscere ed apprezzare un campo di studi del quale non immaginavo nemmeno l'esistenza prima di intraprendere il mio percorso universitario.
Ringrazio Simonetta Montemagni, per avermi seguito e consigliato durante il tirocinio presso l'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR di Pisa, e il successivo lavoro di analisi per questo studio.
Ringrazio Giulia Venturi, e Felice Dell'Orletta, per il sostegno ed il supporto fornitomi.
Ringrazio Francesco, Giulia, Alessandro e Fausto, e tutti gli altri colleghi, amici ed amiche che con me hanno vissuto questa avventura a Informatica Umanistica.
Ringrazio la mia famiglia, che mi ha sostenuto in tutti i sensi, senza la quale non sarebbe stato possibile tutto ciò e in particolare i miei fratelli che con me vivono quest'esperienza universitaria lontano da casa.

1.

Analisi sintattica automatica

1.1 Introduzione su TAL

Lo scopo del Trattamento Automatico del Linguaggio (TAL) consiste nello sviluppo di sistemi che siano in grado di comprendere e/o produrre linguaggio naturale, esattamente come un essere umano.
Il problema è oggi di grande interesse per via dei numerosi risvolti pratici che il TAL può avere:

- estrazione automatica di informazione
- traduzione automatica
- question answering

Questo è ovviamente un compito difficile, a causa dell'ambiguità del linguaggio naturale a tutti i livelli linguistici.
L'approccio al problema è costituito in realtà da più compiti parziali per l'analisi dei diversi livelli linguistici (annotazione morfosintattica, annotazione sintattica, etc.).
Attualmente per creare sistemi che abbiano buone prestazioni in questi compiti, sono utilizzati degli algoritmi di apprendimento automatico supervisionato (machine learning), impiegati per apprendere (con un processo di inferenza induttiva) un modello capace di adempiere al compito, sulla base dei dati d'addestramento annotati.
Per esempio, per l'annotazione morfosintattica i dati di addestramento consistono di documenti annotati con le informazioni lessicali e morfosintattiche di ogni singola parola (lemma, parte del discorso, tratti morfologici).

1.2 Livelli di analisi linguistica

L'analisi linguistica automatica viene normalmente realizzata attraverso una sequenza di fasi di elaborazione su più livelli. A partire dal testo si procede alla divisione in periodi, alla tokenizzazione ed eventualmente alla normalizzazione del testo, alla lemmatizzazione e all'analisi morfosintattica e sintattica, sulla base della quale si può poi procedere all'analisi semantica. Il risultato dell'elaborazione sottostante costituisce di volta in volta l'input per la fase di elaborazione successiva:

1. Divisione in periodi
2. Tokenizzazione: fase di segmentazione del testo in parole-token
3. Normalizzazione: fase di pre-trattamento in cui si cerca di ridurre la variabilità del testo con l'applicazione di alcune procedure standard (ad esempio riduzione degli spazi multipli, aggiunta dello spazio dopo l'apostrofo, trasformazione degli apostrofi in accenti, normalizzazione di numeri, date etc.)
4. Lemmatizzazione: fase in cui le forme flesse sono ricondotte al loro esponente lessicale o lemma
5. Analisi morfosintattica: fase di annotazione delle parti del discorso e dei tratti morfologici

6. Analisi sintattica: fase di annotazione delle relazioni sintattiche a dipendenze (si guardi l'esempio al **paragrafo 1.5**)
7. Analisi semantica: fase di annotazione dell'informazione semantica che si intende rendere esplicita.

1.3 Strumenti e risorse linguistiche per il parsing dell'italiano

1.3.1 Il modulo di annotazione morfosintattica

Il modulo di annotazione morfosintattica sviluppato ed utilizzato, è stato descritto[1] da Dell'Orletta (2009). Con un'accuratezza di circa il 97%, esso si è classificato primo nel Part-Of-Speech Tagging Task della campagna di valutazione Evalita 2009 dimostrandosi in questo modo il più preciso analizzatore morfosintattico (PoS tagger) oggi esistente per la lingua italiana.

1.3.2 Approcci all'annotazione sintattica automatica

Procedendo all'analisi sintattica automatica è possibile applicare principalmente due diverse tipologie di approccio: l'approccio Grammar-Driven (detto anche Rule-based), e l'approccio Data-Driven.
L'approccio **Grammar-Driven** richiede l'utilizzo di grammatiche formali (es. "context-free grammars") ed ha come obbiettivo riconoscere l'appartenenza di una frase specifica a un linguaggio L definito a priori, e renderne esplicita la corretta struttura. L'approccio **Data-Driven** richiede un processo di inferenza induttiva a partire da un corpus "gold standard" annotato manualmente

Il parser sintattico in generale deve tener conto di due principali requisiti:
I.**Robustezza** (capacità di far fronte a situazioni di input mal formato o diverso dal linguaggio per il quale è stato sviluppato).
II.**Disambiguazione** (capacità di disambiguare tra possibili analisi diverse).
L'approccio Data-Driven fa della robustezza la sua caratteristica principale garantendo in ogni caso un'analisi alla frase. Nell'ambito dell'adattamento di dominio, ossia l'adeguamento di un parser ad un dominio linguistico diverso da quello delle risorse linguistiche d'addestramento, è intuitivo che sia la robustezza la caratteristica più utile, ponendo il parser nella situazione di riuscire comunque a proporre un'analisi della frase anche se in contesti critici, ad esempio in frasi di dominio linguistico diverso.

1.3.2.1 DeSR

Il componente di analisi sintattica a dipendenze DeSR (Attardi et al., 2009) è risultato il parser con le migliori prestazioni[2] allo stesso livello del TUP (Turin University Parser)

1 *F. Dell'Orletta.* Ensemble system for part–of–speech tagging. In *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia, 2009.

2 *G. Attardi, F. Dell'Orletta, M. Simi e J. Turian*. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia, 2009.

sviluppato presso l'Università di Torino (Lesmo, 2009). DeSR è il migliore fra gli strumenti Data-Driven, infatti TUP che è uno strumento Grammar-Driven ha mostrato risultati lievemente migliori, anche se la differenza fra le due prestazioni non è statisticamente rilevante (88,73% TUP – 88,67% DeSR).

1.3.3 Corpora rappresentativi della lingua italiana annotati sintatticamente

Uno dei prerequisiti fondamentali per l'uso di metodi Data-Driven è l'utilizzo di grandi quantità di dati testuali opportunamente codificati e annotati al livello linguistico oggetto dell'analisi.
Per l'italiano esistono diversi corpora annotati sintatticamente, in particolare sono da segnalare ISST-TANL (parte di ISST annotata sintatticamente) sviluppato all'ILC del CNR e TUT sviluppato all'università di Torino.

1.3.3.1 ISST-TANL

La Treebank Sintattico-Semantica dell'Italiano (Montemagni et al., 2003) è un corpus della lingua italiana annotato manualmente a più livelli: ortografico, morfo-sintattico, sintattico e lessico-semantico.
Nel contesto del progetto "SI-TAL", finanziato dal Ministero Italiano di Scienza e Ricerca (MURST) per lo sviluppo di una suite integrata di strumenti e risorse per il Trattamento Automatico del Linguaggio per l'italiano, ISST si colloca in posizione di primo piano, rappresentando uno dei principali risultati.
ISST-TANL è un corpus annotato sintatticamente morfo-sintatticamente e sintatticamente a dipendenze, sviluppato, in collaborazione tra l'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR e il Dipartimento di Informatica dell'Università di Pisa, da Montemagni e Simi (2007) partendo dalla revisione di ISST-CoNLL, a sua volta derivato dalla Treebank Sintattico-Semantica Italiana o ISST (Montemagni et al., 2003).
ISST-TANL è costituito da 3.109 frasi, 71.285 tokens[3], e consiste di testi giornalistici e di periodici.
La codifica su ISST-TANL è stata estesa al livello semantico[4] sul modello FrameNet[5] dal Dipartimento di Linguistica dell'Università di Pisa e dall'Istituto di Linguistica Computazionale del CNR (anche grazie al mio lavoro di annotazione manuale tramite lo strumento SALTO[6], durante il mio tirocinio formativo universitario), attualmente su 1.916 frasi per un totale di 2.934 istanze di frame, e 287 frame-tipo annotati.
L'annotazione a dipendenze di ISST-TANL è distribuita su due livelli: morfo-sintattico (parte del discorso e tratti morfologici) e sintattico (relazioni di dipendenza testa-dipendente).
La codifica morfo-sintattica segue il tagset morfo-sintattico di ISST-TANL[7] basato sul

3 *S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi e R. Delmonte. Building and using parsed corpora. In A. Abeill´e, editore, Building and using Parsed Corpora, Language and Speech Series, pp. 189–210. Kluwer, Dordrecht, 2003.*

4 *Alessandro Lenci, Simonetta Montemagni, Giulia Venturi, Maria Rosaria Cutrullà*,"Enriching the ISST–TANL Corpus with Semantic Frames", 2012

5 https://framenet.icsi.berkeley.edu/fndrupal/

6 Erk et *al.*,2003bl

7 http://poesix1.ilc.cnr.it/ISST-TANL-MStagset-web.pdf

tagset ILC/Parole e conforme allo standard internazionale EAGLES, costituito per quanto riguarda la parte del discorso da 14 tag generici (coarse-grained) e 37 tag specifici (fine-grained), e da 6 tratti morfologici obbligatori o opzionali in base ai casi (gen, num, sup, per, mod, tmp)[8].
La codifica sintattica a dipendenze segue il tagset di dipendenza di ISST-TANL[9] che è una nuova versione del tagset ISST-CoNLL.
ISST-CoNLL è stato sviluppato combinando le informazioni provenienti dai livelli di codifica morfo-sintattica e sintattica, producendo semiautomaticamente in output una combinazione delle due informazioni nel formato tabulare CoNLL07.

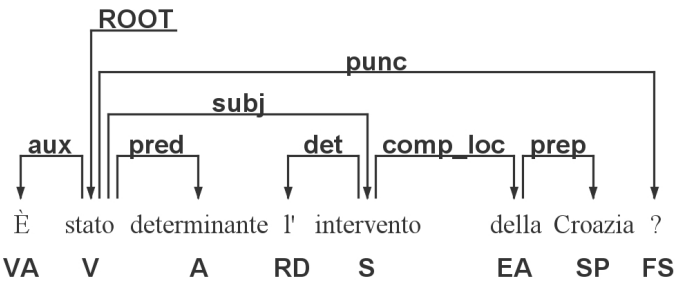
1.4 Rappresentazione a dipendenze: formato CoNLL

Il formato di annotazione a dipendenze utilizzato è conforme al formato tabulare standard CoNLL07, utilizzato in occasione dello "Shared Task on Dependency Parsing" (Nivre et al., 2007).
Ogni parola-token è accompagnata da informazioni sul lemma, sulla parte del discorso sia generica che specifica, sui tratti morfologici, sulla testa e sul tipo della relazione di dipendenza.
L'informazione è distribuita su una riga (una riga per ogni parola nella frase) comprendente otto colonne contenenti le informazioni specifiche. La prima colonna riporta l'informazione sulla posizione del token nella frase, la seconda riporta il token così nella forma specifica con la quale occorre nella frase, la terza riporta il lemma, la quarta riporta la parte del discorso generica, la quinta riporta la parte del discorso specifica, la sesta riporta i tratti morfologici, la settima riporta la posizione della testa sintattica di riferimento, e l'ottava riporta il tipo di relazione sintattica che lega la parola alla testa.
La struttura è quindi di questo tipo:
N- -token- -lemma- -POS gen.- -POS spec.- -tratti- -testa sintattica- -tipo di relazione
Ad esempio la frase "*È stato determinante l'intervento della Croazia?*" è annotata come segue:

1	È	essere	V	VA	num=s|per=3|mod=i|ten=p	2	aux	_	_
2	stato	essere	V	V	num=s|mod=p|gen=m	0	ROOT	_	_
3	determinante	determinante	A	A	num=s|gen=n	2	pred	_	_
4	l'	lo	R	RD	num=s|gen=n	5	det	_	_
5	intervento	intervento	S	S	num=s|gen=m	2	subj	_	_
6	della	di	E	EA	num=s|gen=f	5	comp_loc	_	_
7	Croazia	Croazia	S	SP	_	6	prep	_	_
8	?	?	F	FS	_	2	punc	_	_

Partendo da un'annotazione di questo tipo è possibile implementare una visualizzazione più user-friendly ad albero.
Ad esempio, la visualizzazione in forma di albero a dipendenze della frase in esempio implementata da DgAnnotator[10] (sviluppato da Attardi e Simi presso il Dipartimento di Informatica dell'Università di Pisa) risulterà:

8 http://poesix1.ilc.cnr.it/ISST-TANL-MS_FEATStagset-web.pdf
9 http://poesix1.ilc.cnr.it/ISST-TANL-DEPtagset-web.pdf
10 http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/people.html



La struttura sintattica risulta in questo modo rappresentata tramite archi di dipendenza tra testa sintattiche e dipendenti, in relazioni binarie asimmetriche.
In questo contesto il periodo consiste sostanzialmente di una serie di "nodi lessicali" connessi da archi etichettati con i tipi di dipendenza.

1.5 Misure di valutazione dell'accuratezza

1.5.1 LAS, UAS, LA

Per valutare la validità d'annotazione di un parser, è innanzitutto necessario disporre di risorse linguistiche annotate correttamente.
Questi corpora (generalmente annotati a mano o al più semiautomaticamente, partendo da annotazioni automatiche e procedendo per correzioni manuali) di riferimento sono detti "gold standard" e sono il punto di paragone per quantificare l'accuratezza di analisi dei parser.
Confrontando l'output dell'analisi automatica all'annotazione manuale corretta del gold standard, sarà possibile verificare la corrispondenza dell'analisi all'annotazione esatta. Nel caso particolare dell'annotazione sintattica, un'annotazione può essere completamente corretta, parzialmente corretta, o completamente errata.
È utile in fase di valutazione dell'accuratezza non appiattire le annotazioni parzialmente corrette su quelle completamente errate.
Partendo dall'idea che l'annotazione sia completamente corretta quando nelle triple [testa-dipendente-tipo di relazione] i tre parametri sono corretti (ossia corrispondenti a quelli riscontrati nel gold standatd) è possibile individuare tre misure dell'accuratezza di analisi:

1.*Labelled Attachment Score* (LAS), consiste nella percentuale di tokens cui il parser ha assegnato correttamente sia la testa sintattica sia il tipo di relazione;

2.*Unlabelled Attachment Score* (UAS), consiste nella percentuale di tokens cui il parser ha assegnato correttamente unicamente la testa sintattica;

3.*Label Accuracy score* (LA), consiste nella percentuale di tokens cui il parser ha assegnato correttamente unicamente il tipo di relazione.

Come si può facilmente intuire il LAS è sempre minore o uguale sia dell'UAS che del LA.
In occasione del CoNLL 2007 è stato distribuito lo script eval07.pl utilizzato per il computo di queste tre misure.

1.5.2 Precision, Recall, F-Score

Un'analisi più dettagliata dell'accuratezza dei parser sintattici è conducibile sulla base dei valori percentuali di 'precision' e 'recall' nell'annotazione delle singole relazioni di dipendenza. I valori sono calcolabili in questo modo:

- la 'precision' è data dal rapporto tra il numero di triple [testa, dipendente, tipo di relazione di dipendenza] correttamente individuate (CT: Corrette Trovate) dal parser nel test corpus e il numero totale di triple trovate dal parser, sia corrette che errate (ET: Errate Trovate), nel test corpus;

$$\text{precision} = CT/CT+ET$$

- la 'recall' è data dal rapporto tra il numero di triple [testa, dipendente, tipo di relazione di dipendenza] correttamente individuate dal parser nel test corpus e il numero totale di triple presenti nel 'gold standard', ossia tutte le triple corrette, trovate e non (CN: Corrette Non trovate).

$$\text{recall} = CT/CT+CN$$

Mentre il calcolo della precision permette di valutare la 'precisione' dell'analisi, cioè il numero di analisi corrette rispetto a tutte le analisi realizzate, il valore della recall fornisce indicazioni circa la 'copertura' dell'analisi, il numero cioè di analisi che il parser ha realizzato in modo corretto sul totale di analisi nel 'gold standard' corpus.
Precision e Recall possono combinate in una singola misura F-Score la media armonica di Precision e Recall:

$$\text{f-score} = \frac{2 \times precision \times recall}{precision + recall}$$

Il confronto di tali valori su analisi effettuate su corpora diversi è il punto di partenza per valutare l'accuratezza di un parser su un testo.
Per questo studio sono stati confrontati i valori ottenuti su un corpus della lingua comune e su testi giuridici per verificare l'accuratezza del sistema su testi rappresentativi di questo dominio.

2.

I domini linguistici e la portabilità dei parser: l'adattamento di dominio

2.1 Domini Linguistici e portabilità dei parser

Un problema fondamentale per gli algoritmi di apprendimento automatico è che i sistemi di apprendimento supervisionato dipendono fortemente dai dati sui quali sono stati addestrati.
Ci si aspetta che il modello di analisi sia coerente al massimo con le caratteristiche dei dati di addestramento, perdendo molto in portabilità, e che quindi le prestazioni peggiorino vertiginosamente quando la distribuzione dei dati nel dominio linguistico di addestramento differisce sostanzialmente da quella nel dominio in analisi.
Ad esempio l'accuratezza di un parser statistico addestrato sul Penn Treebank[11] Wall Street Journal (WSJ) diminuisce significativamente quando valutato su testi di domini diversi.
Il "parser Charniak"[12] è una Probabalistic Context-Free Grammar (PCFG) lessicalizzata e fornisce un modello statistico del linguaggio naturale, stimando la probabilità dei vari ruoli sintattici considerati nei dati di training. L'uso della lessicalizzazione e del parse-reranking (Collins, 2000; Charniak e Johnson, 2005) ha portato ad un miglioramento significativo rispetto ai primi modelli non-lessicali.
Il parse-reranking inizia come un parser "standard", ma poi reitera il parsing per generare le n-migliori analisi piuttosto che una singola analisi. Poi una fase di reranking utilizza alcune funzionalità più dettagliate, caratteristiche che è quasi impossibile incorporare nella fase iniziale, per riordinare la lista e scegliere una migliore analisi possibilmente diversa da quella singola che avrebbe elaborato senza il reranking.
Il "parser Charniak", vanta un f-score di 89.7% quando valutato su dati dello stesso dominio dei dati di addestramento (Penn WSJ).
Comunque l'accuratezza si riduce del 6% circa sino all'84.1% sul Brown Corpus che è costituito da testi di varietà maggiore (letteratura).
Precipita ancora sino al 13% in meno (vedi **Tabella 2.1a**), quando è testato su un tipo di linguaggio molto diverso, quale ad esempio il dominio biomedico del GENIA corpus[13], un corpus di abstracts di articoli biomedici, o il parlato delle conversazioni telefoniche dallo Switchboard (SWBD)[14].

Tabella 2.1a

	Test			
Train	WSJ	Brown	Genia	SWBD
WSJ	89,7	84,1	76,2	76,7

Tipo di dati	
WSJ:	articoli di giornale
Brown:	letteratura
Genia	abstracts biomedici
SWBD	conversazioni telefoniche

Tabella 2.1a: valori dell'f-score del parser PCFG Charniak addestrato sul Penn Treebank WSJ e valutato su vari domini (come riportato in McClosky, 2010; p.44)[15]

11 http://www.cis.upenn.edu/~treebank/
12 http://acl.ldc.upenn.edu/P/P06/P06-1043.pdf
13 http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/
14 http://groups.inf.ed.ac.uk/switchboard/index.html
15 http://dissertations.ub.rug.nl/FILES/faculties/arts/2011/b.plank/03c3.pdf

Così il Collins'parser (Collins, 1999) addestrato su testi del Wall Street Journal diminuisce la propria accuratezza di analisi di 5,7 punti percentuali quando testato sul Brown Corpus, passando dall'86,34% all'80,64% di f-score.

Tabella 2.1b

Dati di addestramento	Test	F-Score
WSJ	WSJ	86,34
WSJ	Brown	80,64
Brown	Brown	84,09
WSJ+Brown	Brown	84,33
WSJ+Brown	WSJ	86,59

Tabella 2.1b Risultati del parsing da Gildea[16] (2001) con parser Collins (1999).
Le dimensioni dei training corpora sono rispettivamente:
WSJ 39.832 frasi; Brown 21.818 frasi.

Lo stesso parser, inoltre, come dimostrato da Clegg e Shepherd (2005), addestrato sulla Penn TreeBank (Marcus et al.,1993) e valutato su GENIA, diminuisce di 7,7 punti percentuali, passando dall'86,8% al 79% di precision.
In generale ci sono due approcci principali per affrontare il problema:

1)Annotare manualmente dati per il nuovo dominio

2)Provare ad adattare un modello di parsing di uno specifico dominio d'origine ad un nuovo dominio oggetto di analisi (Adattamento di Dominio).

Annotare manualmente dati per il nuovo dominio è una soluzione costosa e non molto elegante, e quindi, in fin dei conti, non soddisfacente.
Invece l'Adattamento di Dominio cerca di adattare un modello addestrato su un dominio d'origine ad un nuovo dominio oggetto di analisi utilizzando ridotte quantità di dati annotati e/o grandi quantità di dati non annotati del nuovo dominio
In particolare questo lavoro affronta il problema dell'adattamento al dominio giuridico, dominio che si presta perfettamente, data la sua nota complessità e ricercatezza formale, a mettere alla prova i parser sintattici.
L'adattamento proprio a questo particolare dominio è stato infatti oggetto di un task specifico nel contesto della campagna Evalita 2011 (vedi **paragrafo 3.1**).
L'annotazione automatica linguistica di corpora di dominio è un problema affrontato già a partire dagli anni 80, alla ricerca di strategie per adattare gli strumenti all'annotazione di testi di dominio, ma il dominio giuridico in particolare è stato trascurato sino a tempi recentissimi.
Negli ultimi anni sono stati fatti studi che hanno messo in mostra le specializzazioni necessarie all'annotazione di costruzioni linguistiche specifiche del dominio giuridico, e le criticità ai vari livelli di annotazione (vedi **paragrafo 3.2**).

16 http://www.cs.rochester.edu/~gildea/gildea-emnlp01.pdf

2.1.1 Concetto di Dominio

Per affrontare l'argomento è essenziale innanzitutto cercare di definire la nozione di "dominio" in ambito linguistico.
Volendo dare una definizione potremmo dire che un dominio è un "campo", un "ambito", un "settore" della conoscenza (*il dominio dell'arte*, *della scienza, ecc.*).
Nel caso linguistico quindi si tratterà di ambiti linguistici specifici, ovvero di sottolinguaggi.
McClosky (2010) sostiene che il problema della portabilità dei parser "is usually attributed to a difference in domain. By domain we mean the style, genre, and medium of a document.".
Non c'è una base comune che definisca in cosa un dominio consista, in ogni caso termini come "genere", "registro", "tipo di testo", "dominio", "stile", sono spesso usati indifferentemente nelle diverse comunità (D.Y. Lee, 2001).
Intuitivamente i testi variano secondo diversi parametri e caratteristiche.
Ad esempio i testi di giornale contengono frasi relativamente lunghe tipiche del registro scritto (ciò vale ancor più per il dominio che prenderò in esame, ossia il dominio giuridico).
Al contrario i dati costituiti da trascrizioni del parlato contengono molte idiosincrasie del linguaggio parlato, quali esitazioni, ripetizioni, false partenze, forme contratte, segnali extra-linguistici (pause).
Anche il linguaggio scritto sui recenti social network contiene molte contrazioni, ma inoltre contiene smiles, errori di scrittura, token particolari quali le URL o gli username.
Questo tipo di linguaggio può essere considerato addirittura più distante dallo scritto rispetto al parlato.
Un semplice ed ovvio parametro variabile fra i vari domini è la lunghezza delle frasi.
Anche solo intuitivamente, risulta chiaro che la lunghezza media delle frasi in trascrizioni del parlato sia piuttosto differente (in particolare minore) da quella delle frasi in un quotidiano.
Le trascrizioni del parlato contengono una grande quantità di frasi molto brevi.
Anche corpora rappresentativi (almeno negli intenti) dello stesso dominio, possono avere differenze di questo tipo, ad esempio la lunghezza media delle frasi nel Brown è minore di quella nel WSJ.
Invece la lunghezza media delle frasi del WSJ e di Genia sono relativamente simili, osservando solo la forma e la distribuzione, ed ignorando la frequenza, non saremmo nemmeno in grado di distinguerli.
È chiaro che ci sia più di un parametro di misurazione delle differenze fra testi.
Genia contiene un gran numero di termini altamente specializzati, quindi le differenze di vocabolario possono essere una misura considerevole della differenza di dominio tra Genia e WJS.
La differenza fra due domini può quindi essere espressa principalmente anche da un solo parametro, diverso in base alla situazione; di conseguenza, anche le diverse strategie di adattamento possono essere più o meno efficaci, a seconda della corrispondenza o meno della strategia scelta al parametro di differenza maggiormente significativo fra i due domini in questione.
Ad esempio per l'adattamento al dominio biomedico buona parte del miglioramento sarà da attribuire al riaddestramento del PoS tagger, mentre per l'adattamento al dominio del parlato, ricco di proposizioni interrogative sarà da attribuire al riaddestramento del parser

sintattico.
Intuitivamente, infatti, la differenza principale fra il linguaggio giornalistico ed il linguaggio biomedico è lessicale, mentre la differenza principale fra il linguaggio giornalistico e il linguaggio parlato ricco di proposizioni interrogative è sintattica.
In alcuni casi risulterà meno chiaro quale parametro sia maggiormente significativo, ad esempio volendo analizzare le differenze fra il dominio del parlato, ricco proposizioni interrogative e il dominio biomedico noteremmo una grande differenza fra i due tipi di testo.
Sempre poco chiaro sarebbe individuare il parametro più significativo nel caso in cui dovessimo adattare un parser addestrato su testi giornalistici ad un dominio "biomedico parlato", ricco di proposizioni "interrogative biomediche".
In generale sia le differenze sintattiche, sia quelle lessicali sono fondamentali.
Tornando al concetto di dominio in generale in passato il termine è stato usato arbitrariamente per riferirsi ad un qualche tipo di unità linguistica coerente (per argomento, stile, registro o genere).
Volendo porre la questione nel caso limite "each document is potentially its own domain" (McClosky et al. 2010).
Si potrebbe supporre un'assunzione implicita: che ogni dominio sia rappresentato dal rispettivo corpus, considerando un corpus come un'unità omogenea.
In realtà un corpus può contenere una grande varietà di generi, quindi l'idea che sia il corpus stesso a costituire il dominio non è affatto[17] una teoria universalmente condivisa.
Barbara Plank e Gertjan van Noord (2011) suggeriscono di spezzettare il corpus a livello di articoli.
È infatti più verosimile che singoli articoli siano espressione di un unico dominio linguistico.
La questione non è quindi di facile soluzione ed è ancora oggetto di ricerca. Sempre Barbara Plank e Gertjan van Noord (2011) cercando di definire una misura di similarità fra domini, si chiedono se questa misura, una volta definita compiutamente non possa dirci qualcosa di più riguardo cosa si intenda per "dominio".

2.1.2 I Primi studi sull'Adattamento di Dominio

Per lo studio e la ricerca in quest'ambito è normale partire da Gildea 2001, che giustamente si chiede:
"*Can training data from one corpus be applied to parsing another?*".
I primi studi sull'Adattamento di Dominio e in generale sulla portabilità dei parser risalgono a Sekine (1997), Ratnaparkhi (1999), e Gildea (2001).
Sekine (1997) sostiene che "it is intuitively conceivable that there are syntactic differences between 'telegraphic messages' and 'press reports'[...]".
Nello studio dell'adattamento di dominio, come fatto osservare da Clegg e Shepherd (2005), è opportuno cominciare dall'analisi degli errori per identificare le fonti di problemi, fase centrale per lo sviluppo di una strategia di adattamento degli strumenti.
Clegg e Shepherd stati i primi a intraprendere studi empirici comparativi delle distribuzioni nei sottodomini del Brown Corpus.
Mentre Sekine (1997) ha analizzato le performance dei parser con i sottodomini del Brown corpus, Ratnaparkhi (1999) e Gildea (2001), sono stati i primi ad esaminare l'accuratezza

17 Sekine, 1997; Kilgarriff e Rose, 1998; Plank e Sima'an, 2008; Webber, 2009; Lippincott, O' Séaghdha, Sun e Korhonen, 2010

dei parser su diversi corpora (tra WSJ e Brown).
Entrambi hanno analizzato le performance di un parser addestrato sul WSJ e testato sul Brown corpus, rilevando un calo nell'accuratezza di circa il 6%.
Applicando una semplice combinazione dei dati (WSJ+Brown) è stato ottenuto un miglioramento di accuratezza molto piccolo (0.25%) (vedi **Tabella 2.1b**).
I risultati di questi studi portarono alla conclusione che una grande quantità di dati Fuori-Dominio non aiuta, anzi, una piccola quantità di dati d'addestramento annotati sembra essere più utile di una grande quantità di dati non annotati (Gildea 2001).
Lease e Charniak (2005) sono stati i primi ad analizzare la portabilità di un parser ad un altro dominio, in particolare a quello biomedico (Genia). Hanno mostrato che la percentuale di parole sconosciute aumenta passando a domini tecnici e ciò chiaramente influisce sull'accuratezza del parsing. Gli studi più recenti hanno definito l'approccio Data-Driven come stato dell'arte per l'adattamento di dominio (Plank e van Noord 2010).
Il limite di questo approccio è il fatto di essere legato ad un training corpus di dominio manualmente annotato.

2.1.3 Caratteristiche del parser e portabilità, sensibilità alla variazione di dominio: Grammar-Driven versus Data Driven

Il problema della dipendenza dal dominio costituisce una sfida per entrambi gli approcci al parsing Grammar-Driven e Data Driven. I due approcci hanno in comune l'inadeguata portabilità ai nuovi domini. Le prestazioni crollano significativamente all'aumentare della differenza tra il dominio nuovo e il dominio di addestramento. Ma quale dei due tipi di parsing soffre maggiormente questo tipo di problemi? È probabilmente intuitivo che un parser scritto a mano di tipo grammar-driven sia pensato per ottenere la copertura più estesa possibile. Barbara Plank ha investigato quest'ipotesi nella sua Tesi di Dottorato ponendosi la domanda nei termini specifici:

"*Which parsing system is more robust with respect to different input texts?*"

Il suo studio ha messo a confronto le accuratezze di analisi di diversi parser per l'olandese: Alpino (Grammar-Driven), MST (Data-Driven) e Malt (Data-Driven), arrivando alla conclusione, nel caso del suo esperimento che in generale, i sistemi Data-Driven si affidano pesantemente sui dati di addestramento per calcolare i loro modelli.
Aggiungendo però che questo non significa affatto che i sistemi Grammar-Driven non soffrano questo problema. Come riscontrato da Zhang e Wang (2009), la componente di disambiguazione e la copertura lessicale dei sistemi Grammar-Driven sono comunque dipendenti dal dominio. Quindi la dipendenza dal dominio rimane un problema per entrambi i tipi di parser. I risultati della sua ricerca mostrano in ogni caso una maggiore robustezza del parser Grammar-Driven, ma sono comunque risultati specifici per l'olandese, e sarebbe interessante applicare questo tipo di valutazioni anche ad altre lingue e ad altri sistemi.

2.2 Adattamento di dominio su strumenti Grammar-Driven

L'adattamento di dominio su strumenti Grammar-Driven è molto costoso e consiste nell'estensione manuale della grammatica funzionale. Questo approccio, per quanto consenta una maggiore indipendenza dai dati linguistici annotati è allo stato attuale poco economico richiedendo un intervento massiccio di estensione per l'adattamento a domini specifici. In ogni caso, solo pochi studi hanno esaminato (e.g. Hara et al., 2005; Plank, 2009b) il problema provando ad adattare la componente di disambiguazione di un sistema grammar-driven.

2.3 Adattamento di dominio su strumenti Data-Driven

Esistono tre principali approcci Data Driven all'Adattamento di Dominio:
1.Adattamento di dominio supervisionato (e.g. Hara et al., 2005; Daumé III, 2007)
2.Adattamento di dominio non supervisionato (e.g. Blitzer, McDonald e Pereira, 2006; McClosky et al., 2006)
3.Adattamento di dominio semi-supervisionato (e.g. Daumé III, Kumar e Saha, 2010; Chang, Connor e Roth, 2010)
La scelta di uno dei diversi approcci è legata al tipo di dati a disposizione per il nuovo dominio (dati annotati, dati non annotati, sia dati annotati che dati non annotati). Questo approccio rimane più dipendente dalle risorse linguistiche annotate rispetto a quello Grammar-Driven, ma è più economico.

2.3.1 Adattamento di Dominio Supervisionato

L'Adattamento di dominio supervisionato (e.g. Hara et al., 2005; Daumè III, 2007) è lo scenario in cui abbiamo accesso ad una grande quantità di dati annotati del dominio di addestramento originario (dati sorgente) e solo ad una quantità comunque limitata di dati annotati del nuovo dominio di destinazione. Quindi, dato un sistema di analisi e di due insiemi di dati, i dati sorgente del dominio originale (Fuori-dominio) e i dati di destinazione (nuovi o In-dominio), l'obiettivo è quello di adattare il sistema di analisi (che è di default addestrato sui dati fuori-di-dominio) al nuovo dominio di destinazione.
In particolare disponiamo di una grande quantità di dati sorgente annotati ed una ridotta quantità di dati annotati del nuovo dominio.

2.3.2 Adattamento di Dominio Non Supervisionato

Nell'Adattamento di Dominio Non Supervisionato invece di disporre di dati di addestramento del nuovo dominio annotati, si dispone solo di dati non annotati.
È molto facile disporre di dati non annotati ma l'adattamento di dominio non supervisionato è considerevolmente più difficile.

2.3.3 Adattamento di Dominio Semi-supervisionato

Solo studi recenti hanno cominciato a utilizzare piccole quantità dati del nuovo dominio annotati insieme a nuovi dati non annotati.

2.3.4 Addestramento In-Dominio, Fuori-Dominio, Combinato

Per l'Adattamento di Dominio esistono diversi approcci:
1)Fuori-Dominio
2)In-Dominio
3)Combinato
L'approccio più ovvio è quella Fuori-Dominio: consiste nell'addestrare un modello sui dati del dominio di origine, ignorando il tipo di dati oggetto dell'analisi.
Questo è il modello predefinito prima di applicare qualunque tecnica di adattamento.
Purtroppo le performance del modello Fuori-Dominio sono compromesse quando esso è utilizzato su dati di un nuovo dominio.
L'Adattamento di Dominio intende potenziare questo tipo di modello.
Con l'Adattamento di Dominio Supervisionato sono applicabili sia l'approccio In-Dominio, sia l'approccio Combinato.
Nel primo caso abbiamo un modello addestrato solo su dati In-Dominio.
Nel secondo caso, si cerca di superare i limiti posti dalla ridotta quantità di dati In-Dominio annotati, sfruttando la maggiore quantità di dati Fuori-Dominio, unendo dati In-Dominio e dati Fuori-Dominio. Questo modello viene definito "Combinato".
Questa soluzione richiede un tempo di addestramento considerevolmente maggiore, inoltre la preponderante quantità di dati Fuori-Dominio potrebbe sommergere la molto ridotta quantità di dati In-Dominio.
Per ovviare al problema è stata ideata una variante che bilanci le differenti quantità di dati "pesando" adeguatamente i dati di addestramento.
Un modo per superare le differenze tra In-Dominio e Fuori-Dominio è incorporare il modello del dominio d'origine come punto di partenza nella creazione di un modello per il nuovo dominio per il quale si disponga di risorse linguistiche limitate.
Già in passato Hara et al. (2005) hanno tentato di adattare un parser addestrato sul WSJ al dominio biomedico (GENIA), integrando il modello originale (grande quantità di dati Fuori-Dominio) come distribuzione di riferimento nell'addestramento di un modello per il nuovo dominio.
L'idea è di sfruttare le informazioni del modello più generale, stimato dalla più grande, Treebank Fuori-Dominio, per analizzare dati di un particolare dominio per il quale è disponibile solo una piccola quantità di dati di addestramento.
I loro risultati (vedi **Tabella 2.3.4**) mostrano come incorporare il modello generale abbia aiutato nell'incrementare le performance dell'approccio In-Dominio. In ogni caso la semplice combinazione dei dati (cioè l'unione dei corpora In-Dominio e Fuori-Dominio di addestramento) ha ottenuto risultati molto simili a quelli ottenuti con il loro metodo. A prescindere dal metodo di integrare il modello originale come distribuzione di riferimento, risulta chiaro che l'apporto di una grande quantità di dati Fuori-Dominio contribuisce significativamente a migliorare le prestazioni del parser addestrato sui soli dati In-Dominio.

Tabella 2.3.4	**F-Score**
Reference distribution (Hara et al., 2005)	86,87
Combinazione dei dati	86,32
Solo Genia (baseline In-dominio)	85,72
WSJ, modello originale (Fuori-dominio)	85,10

Risultati riportati in Hara et al. (2005), dove il modello generale è stato impiegato come distribuzione di riferimento per il dominio oggetto di analisi. Le dimensioni di Genia e WSJ sono, rispettivamente: 3.524 e 39.832 frasi.

2.4 Misura della similarità fra domini

Oltre ad aiutarci a definire il concetto di "dominio", una misura della similarità fra domini potrebbe contribuire in maniera significativa al processo di Adattamento di Dominio rendendo esplicito quali dati o modelli siano i più idonei a migliorare le performance del parsing su un nuovo dominio.
Van Asch e Daelemans (2010) hanno studiato una misura di differenza tra domini e la correlazione con l'accuratezza dell'annotazione morfosintattica automatica.
Lippincott et al. (2010) hanno esaminato variazioni di sotto-dominio nei corpora biomedici.
McClosky et al. (2010) hanno provato a trovare la migliore combinazione di modelli di addestramento per analizzare dati di un nuovo dominio.
Hanno usato dati non annotati per creare alcuni parser pesati sulla base della similarità al sotto-dominio.
Hanno addestrato un modello di regressione lineare per prevedere i migliori (interpolazione lineare) modelli del dominio d'origine. McClosky et al. (2010) hanno considerato il nuovo dominio come un misto di domini d'origine.
Plank et van Noord hanno fatto lo stesso considerando gli articoli come unità base e hanno provato a trovare sottoinsiemi di articoli correlati direttamente applicando misure di similarità tra domini.
Una funzione di similarità può essere definita su alcuni insiemi di eventi considerati rilevanti, quali: parole, caratteri, ngrammi (di parole o caratteri), annotazioni morfosintattiche, dipendenze lessicali, ruoli sintattici.
Sono state ipotizzate misure della similarità fra domini di vario tipo, principalmente appartenenti a due categorie: funzioni motivate probabilisticamente (ad esempio la divergenza di Kullback-Leibler, di Jensen-Shannon) e funzioni motivate geometricamente (la distanza Euclidea, la misura del coseno dell'angolo fra due vettori, e la distanza variazionale).

3.

Adattamento al dominio giuridico: "Domain Adaptation" task in Evalita 2011

3.1 "Domain Adaptation" task in Evalita 2011

Il gruppo di lavoro per il Trattamento Automatico del Linguaggio sostenuto da AI*IA[18] (Associazione Italiana per l'Intelligenza Artificiale) e AISV[19] (associazione italiana di Scienze della Voce), dopo il successo di Evalita 2007 ed Evalita 2009, ha organizzato Evalita 2011[20], la terza campagna di valutazione degli strumenti di Trattamento Automatico del Linguaggio e degli strumenti vocali per l'italiano.
I compiti del progetto sono divisi sostanzialmente in “Text tasks” (per quanto riguarda il Trattamento Automatico del Linguaggio testuale) e “Speech tasks” (per quanto riguarda gli strumenti vocali automatici per l'italiano).
Per quanto riguarda questo lavoro, è di nostro interesse il Trattamento Automatico del Linguaggio testuale, ed in particolare il task di “Domain Adaptation”.
Questo task in particolare è stato seguito dall'Istituto di Linguistica Computazionale del CNR di Pisa[21] e dall'Istituto di Teoria e Tecniche dell'Informmazione Giuridica del CNR di Firenze, in particolare da Felice dell'Orletta (ILC-CNR, Pisa), Simonetta Montemagni (ILC-CNR, Pisa), Giulia Venturi (ILC-CNR, Pisa), Tommaso Agnoloni (ITTIG-CNR, Firenze), Enrico Francesconi (ITTIG-CNR, Firenze), e Simone Marchi (ILC-CNR, Pisa).
È nell'ambito di questo task che si inserisce questo lavoro.
Il task di adattamento di dominio mira a studiare e definire le tecniche allo stato dell'arte per l'adeguamento dei sistemi di analisi sintattica a dipendenza ai domini linguistici diversi da quello dei dati su cui sono stati addestrati o sviluppati. Questa è la prima volta che tale compito viene proposto nell'ambito della campagna EVALITA e per la lingua italiana.
Nell'ambito del tirocinio formativo universitario previsto dal piano di studi del corso di laurea in Informatica Umanistica, ho partecipato, presso l'ILC del CNR, al lavoro, tramite l'uso del tool DgAnnotator[22], di annotazione semi-automatica (ossia realizzata tramite correzioni di annotazioni precedentemente elaborate in maniera automatica dal parser) per l'estensione del corpus, TEMIS (vedi **paragrafo 3.3.1**) insistendo sull'osservazione dei tratti specifici di dominio già rilevati da Venturi (2011), allo scopo di annotare a mano (in fase di correzione) in maniera da conseguire il migliore risultato.
Ho poi preso parte anche al lavoro, inserito e descritto in SPLeT 2012[23], di annotazione semantica su TEMIS effettuato manualmente (su modello FrameNet) introducendo frame deontici (tramite lo strumento SALTO[24]).
Al termine del lavoro realizzato nel contesto del tirocinio formativo universitario presso l'ILC del CNR, ho intrapreso uno studio dei dati annotati dal parser in alcune diverse situazioni di analisi (i corpora TEMIS e ISST-TANL usati in diverse combinazioni come

18 https://sites.google.com/a/aixia.it/nlp/
19 http://www.aisv.it/index.php
20 http://www.evalita.it/2011/
21 poesix1.ilc.cnr.it/evalita2011/
22 http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/people.html
23 *Giulia Venturi* (2012), Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts, in Proceedings of the 4th Workshop "Semantic Processing of Legal Texts" (SPLeT 2012), held in conjunction with LREC 2012, Istanbul, Turkey, 27th May, 2012.
24 Erk et al.,2003bl

training per il parser DeSR testato su target corpora estratti anch'essi da TEMIS, nei vari sotto-domini, e da ISST-TANL).
In quest'ambito per l'addestramento del parser il punto di partenza è stato ISST-TANL.
Si procederà a testare l'adattamento al dominio giuridico, tramite l'utilizzo di target corpora giuridici costituiti da leggi emanate da differenti Istituzioni legislative (la Commissione Europea, il Parlamento Italiano e le Regioni), che regolano vari settori, dalla lotta all'inquinamento, ai diritti umani, ai diritti dei disabili, alla libertà d'espressione.
Questo studio dei dati annotati dal parser nelle diverse combinazioni di training e di target, è l'oggetto di questo lavoro.

3.2 I tratti "problematici" dell'italiano giuridico

La maggior parte dei tratti critici dell'italiano giuridico sono già stati analizzati ed esposti da Venturi (2011), essi sono:

i.Lunghezza delle relazioni di dipendenza: le frasi del dominio giuridico sono generalmente più lunghe rispetto a quelle dell'italiano standard, ad esempio nel caso dei corpora oggetto di questo lavoro la lunghezza media delle frasi per l'italiano giornalistico è di circa 20 parole, mentre per i diversi corpora giuridici oscilla tra le 30 e le 40 parole. Ciò comporta la possibilità di lunghe relazioni di dipendenza, in casi limite anche di diverse decine di parole.

ii.Usi di parole in contesti specifici di dominio diversi da quelli propri della lingua comune (ad esempio "direttiva", "data", "allegato", prevalentemente participi nella lingua comune, sono principalmente sostantivi nella lingua del diritto) .

iii.Costruzioni sintattiche particolari, quali costruzioni ellittiche (durante il mio lavoro di correzione delle annotazioni automatiche ho avuto largo riscontro di queste ultime), frasi participiali e sequenze di strutture appositive quali partizioni interne dell'articolato di un atto normativo (ad esempio "di cui all'articolo 30", "considerato il parere tecnico...", "fatto salvo il diritto...", "la legge 130, titolo I, articolo 4, comma 5,...").

iv.Densità dei vizi interpuntivi: il dominio giuridico fa largo uso di elenchi di disposizioni separate da virgole o da punti e virgola, spesso introdotti dai due punti. Un uso della punteggiatura al quale il parser addestrato sull'italiano giornalistico non è preparato.

In fase di correzione manuale delle annotazioni automatiche elaborate dal parser, ho riscontrato alcuni casi particolari d'interesse:

1. Prendere atto:

“...

2	prendere	prendere	V	V	mod=f	1	prep	_	_
3	atto	atto	S	S	num=s|gen=m	2	obj	_	_
4	,	,	F	FF	_	5	punc	_	_
5	per	per	E	E	_	1	comp	_	_
6	le	il	R	RD	num=p|gen=f	7	det	_	_
7	ragioni	ragione	S	S	num=p|gen=f	5	prep	_	_
8	in	in	E	E	_	10	comp	_	_
9	premessa	premessa	S	S	num=s|gen=f	8	prep	_	_
10	indicate	indicare	V	V	num=p|mod=p|gen=f	7	mod	_	_
11	,	,	F	FF	_	5	punc	_	_
12	delle	di	E	EA	num=p|gen=f	2	comp	_	_
13	attività	attività	S	S	num=n|gen=f	12	prep	_	_

...”

In fase di correzione delle annotazioni si è considerato “atto”(3) *obj* di “prendere”(2), anche se "prendere atto" è una costruzione fissa che andrebbe probabilmente considerata un tutt'uno. Questa è una peculiarità del dominio giuridico, più o meno raramente riscontrabile nella lingua italiana comune.

2. Variazione nella realizzazione sintattica di “modificatori” coordinati a pari livello di profondità nell'albero di dipendenza:

“...

83	violazioni	violazione	S	S	num=p|gen=f	78	disj	_	_
84	di	di	E	E	_	83	comp	_	_
85	leggi	legge	S	S	num=p|gen=f	84	prep	_	_
86	di	di	E	E	_	85	mod	_	_
87	parità	parità	S	S	num=n|gen=f	86	prep	_	_
88	o	o	C	CC	_	86	dis	_	_
89	comunque	comunque	B	B	_	90	mod	_	_
90	attinenti	attinente	A	A	num=p|gen=n	86	disj	_	_
91	alla	al	E	EA	num=s|gen=f	90	comp	_	_
92	condizione	condizione	S	S	num=s|gen=f	91	prep	_	_
93	della	di	E	EA	num=s|gen=f	92	comp	_	_
94	donna	donna	S	S	num=s|gen=f	93	prep	_	_

...”

In questo caso abbiamo un elenco di modificatori coordinati da congiunzioni disgiuntive “o”. Il primo modificatore, in questo caso “di parità”, si realizza come complemento di specificazione, mentre il secondo si realizza come attributo e quindi “modificatore” in senso stretto. In altri casi può succedere il contrario e l'elenco può anche contenere più di due diversi “modificatori” realizzati in maniera sintatticamente differente.
In un caso di medesima realizzazione sintattica dei diversi modificatori elencati, il secondo modificatore e gli eventuali modificatori successivi nell'elenco sarebbero stati annotatati come *disj* (o *conj* nel caso di congiunzioni non disgiuntive) del primo modificatore dell'elenco, in questo caso“di”(86) di "di parità", e le congiunzioni interposte sarebbero state annotate come *dis* (o *con*), sempre del primo modificatore “di”(86).
Solo il primo modificatore sarebbe stato annotato con la relazione sintattica diretta con la testa (in questo caso *comp*) “leggi”(85). Nell'esempio presentato l'annotazione tenta di

seguire l'approccio “tipico” nonostante questo caso sia “anomalo”, per mostrare in che modo l'annotazione rappresenti generalmente le strutture di coordinazione.

Ulteriori tratti possono essere misurati eseguendo il cosiddetto monitoraggio linguistico tramite gli strumenti di trattamento automatico del linguaggio.
È stato infatti dimostrato da Dell'Orletta e Montemagni (2010a) che dall'annotazione è possibile ricavare informazioni utili alla ricostruzione del profilo linguistico di un testo. È una metodologia affidabile per condurre indagini quantitative del profilo linguistico di testi giuridici e per la definizione di basi teoriche per lo sviluppo di strumenti di monitoraggio della redazione di atti chiari, semplici e comprensibili basati su indicatori di leggibilità.
È possibile restituire il profilo linguistico dei testi e osservare che similarità e differenze ai vari livelli di annotazione corrispondono a tratti specifici.
Ad esempio si può osservare a livello generale la lunghezza media dei tokens e la lunghezza media dei periodi, a livello morfosintattico la distribuzione delle varie parti del discorso (participi, verbi, sostantivi) e delle persone verbali, a livello sintattico la distribuzione dei tipi di relazione di dipendenza, la lunghezza media delle relazioni, il livello di incassamento gerarchico e le dipendenze di predicati verbali, a livello lessicale la densità lessicale (lessico referenziale/lessico funzionale), e la ricchezza lessicale (Type-Token-Ratio).

3.3 Risorse Linguistiche per il dominio giuridico

3.3.1 TEMIS

Tabella 3.3.1A

Corpus	No. tokens	No. sentences
TEMIS-EU-GOLD	6683	275
TEMIS-NAT-GOLD	3670	94
TEMIS-LOC-GOLD	5453	135
TEMIS-GOLD	15804	504

Il corpus TEMIS (Syn**T**actically and S**EM**antically Annotated **I**talian Legislative Corpu**S**) è stato creato in prima istanza nel quadro della tesi di dottorato di Venturi (2011), e poi ingrandito[25](come già scritto, anche grazie alla mia partecipazione all'annotazione semi-automatica nel contesto del mio tirocinio formativo universitario) nel corso della campagna Evalita 2011 durante la quale è stato usato un sottoinsieme di esso per il task di “Domain Adaptation”.
TEMIS è una raccolta di testi legislativi emanati da tre diverse istituzioni, cioè la Commissione Europea (sub-varietà TEMIS-EU), lo Stato Italiano (sub-varietà TEMIS-NAT) e la Regione Piemonte (sub-varietà TEMIS-LOC) (vedi **Tabella 3.3.1A**), che regolano una varietà di argomenti, che vanno dall'ambiente, ai diritti umani, ai diritti delle persone disabili alla libertà di espressione. Si tratta di una raccolta di documenti eterogenei (vedi **Tabella 3.3.1B**), inclusi gli atti giuridici come leggi nazionali e regionali, le direttive europee, decreti legislativi, ecc, così come gli atti amministrativi, quali circolari ministeriali, ecc.

25 *Giulia Venturi* (2012), Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts, in Proceedings of the 4th Workshop "Semantic Processing of Legal Texts" (SPLeT 2012), held in conjunction with LREC 2012, Istanbul, Turkey, 27th May, 2012.

La logica alla base della costruzione di un tale corpus eterogeneo è di raccogliere testi di legge che esemplifichino diverse sotto-varietà della lingua giuridica italiana.

Tabella 3.3.1B	
Corpus	Lunghezza media delle frasi (in tokens)
ISST-TANL	21,87
TEMIS	31,36
TEMIS-EU	24,56
TEMIS-NAT	39,04

3.3.2 ISST-TANL e TEMIS a confronto

Per ottenere la prova della specificità linguistica dei testi legislativi inclusi in TEMIS, il corpus è stato esaminato rispetto ad un numero di parametri diversi, che secondo la letteratura sulla variazione di registro (Biber e Conrad, 2009) sono indicativi delle differenze di genere testuale. Si va dalle caratteristiche primarie del testo, come la lunghezza della frase, a quelli più complessi (ad esempio l'altezza dell'albero sintattico) rilevati dal livello sintattico dell'annotazione.
Un confronto con le rispettive caratteristiche corpus di un quotidiano italiano, scelto per essere rappresentativo della Lingua italiana comune, aiuta a evidenziare le principali caratteristiche linguistiche dei testi legislativi di Temis.
A tale scopo è stato utilizzato l'ISST-TANL corpus, sviluppato congiuntamente dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) e dall'Università degli Studi di Pisa nel quadro del progetto TANL (Analisi del Testo e elaborazione del Linguaggio Naturale).
I corpora TEMIS e ISST-TANL differiscono in modo significativo in molti aspetti a partire dalla lunghezza media della frase, calcolata come il numero medio di parole per frase. Alcune differenze possono anche essere trovate tra i tre tipi di testi giuridici considerati sub-varietà di dominio. In particolare, i testi di legge emanata dalla Commissione Europea (TEMIS-EU) mostrano un comportamento che è più simile al linguaggio ordinario rispetto che ai testi di legge nazionale (TEMIS-NAT) e locale (TEMIS-LOC) .
È interessante notare che varie differenze si possono trovare tra i due corpora rispetto alla distribuzione delle caratteristiche tipicamente correlate alla complessità del testo, come ad esempio la profondità dell'albero di analisi sintattica e la lunghezza dei legami di dipendenza. Secondo l'approccio al monitoraggio linguistico descritto in (Dell'Orletta et al., 2011), i corpora TEMIS e ISST-TANL sono stati confrontati rispetto rispetto a:

i)la lunghezza media dei legami di dipendenza, misurati in termini di parole che occorrono tra la testa sintattica e il dipendente

ii)la profondità media dell'albero sintattico, calcolata in termini di percorso più lungo dalla radice dell'albero delle dipendenze in una certa foglia.

I risultati del confronto, rivelano che le frasi legislative contengono dei legami di dipendenza (14,5) molto più lunghi in media rispetto a quelli delle frasi dell'italiano comune (8,61) e che ii) l'altezza media degli alberi di analisi sintattica di TEMIS (7,44) è superiore a quella che caratterizza le frasi di ISST-TANL (5,28). Inoltre i testi giuridici

europei hanno caratteristiche sintattiche che li rendono più simili al linguaggio ordinario che ai testi nazionali e locali. A questo proposito vale la pena notare che le frasi di TEMIS sono caratterizzate da una profondità media di 'catene' di complementi "a cascata" governate da una testa nominale contenenti sia complementi preposizionali sia nominali e modificatori aggettivali (1,54) superiore a quella delle frasi di ISST-TANL (1,28). Tuttavia, la caratteristica fondamentale distintiva di frasi legislative sembra essere la distribuzione percentuale diversa delle profondità delle 'catene' di complementi "a cascata". I testi legislativi sembrano avere una percentuale maggiore di 'catene' di complementi profonde rispetto al corpus di riferimento italiano.

3.3.3 Test Corpora

Per valutare l'accuratezza del parser sintattico sono stati utilizzati dei test-corpora annotati manualmente chiamati convenzionalmente "gold": ISST-TANL-GOLD, TEMIS-GOLD, TEMIS-EU-GOLD, TEMIS-NAT-GOLD, TEMIS-LOC-GOLD. Per quanto riguarda il dominio dell'italiano comune il test-corpus utilizzato è ISST-TANL-GOLD (da qui in poi "ISST") (5165 token, 231 frasi), mentre per quanto riguarda il dominio giuridico è stato utilizzato TEMIS-GOLD (da qui in poi "TEMIS") (5866 token, 192 frasi), costituito da: TEMIS-EU-GOLD (da qui in poi "TEMIS-EU") (1932 token, 96 frasi), TEMIS-NAT-GOLD (da qui in poi "TEMIS-NAT") (1971 token, 37 frasi), TEMIS-LOC-GOLD (da qui in poi "TEMIS-LOC") (1963 token, 59 frasi).

3.3.4 Estensione dei criteri di annotazione per il dominio specifico

I normali criteri di annotazione utilizzati per ISST, nell'ambito del task di Domain Adaptation, sono dovuti essere estesi (secondo alcuni criteri già individuati in Venturi 2011) per rispondere alle peculiarità del dominio giuridico in parte descritte al **paragrafo 3.2**. Le estensioni sono relative a diversi livelli di annotazione del testo, dalla divisione in frasi all'annotazione a dipendenze. La divisione in frasi, ad esempio, intende in questo caso preservare la struttura originaria del testo giuridico. Sono frequenti nei testi giuridici alcune particolari situazioni di interpunzione. Nei preamboli legislativi, ricorrono frasi che cominciano con costrutti del tipo "Considerato che" e finiscono con segni di interpunzione del tipo ";", frasi che finiscono con segni di interpunzione del tipo ":" ed introducono una lista di elementi e frasi, elementi di una lista, che terminano con segni di interpunzione del tipo ";". Un estensione a questo livello consiste nel trattare i segni di interpunzione del tipo ";" e ":" seguiti da ritorno a capo come segni di fine frase. Estensioni di questo tipo consentono di tenere conto di casi specifici frequenti nei testi giuridici. Anche al livello dell'annotazione a dipendenze è stato necessario apportare delle estensioni (Venturi 2011) per risolvere alcuni casi principali:

1. Costruzioni ellittiche frequentemente usate nelle citazioni a testi giuridici o a sottosezioni specifiche di essi (paragrafo, articolo, comma). Questo è il caso, ad esempio, della frase "*in base ai criteri di cui al paragrafo 7, le misure sono obbligatorie nelle zone seguenti*:". Poiché manca il verbo nella relativa "*di cui al paragrafo 7*" si individua una dipendenza di modificatore relativo fra "*criteri*" che è la testa sintattica e "*al*" che è il token dipendente.

2. Frasi participiali del tipo "*fatto salvo*", utilizzate per esprimere eccezioni o limitazioni. Ad esempio in "*La presente convenzione si applica anche al trasporto di cui al capitolo V, fatte salve le disposizioni ivi previste.*" è stata individuata una dipendenza di modificatore fra la testa della frase participiale ("*fatte*") e la testa sintattica della frase principale ("*applica*").

3. Grandi distanze testa-dipendente, molto frequenti in questo dominio, causa di archi non proiettivi (archi testa-dipendente che si incrociano con altri archi di dipendenza). Mc Donald e Nivre (2007) hanno osservato che DeSR ha una diminuzione di prestazioni nel riconoscimento di periodi molto lunghi, dovuta alla distanza tra un dipendente e la sua testa sintattica che, aumentando il numero di possibili scelte genera ambiguità di analisi. È da notare anche la massiccia presenza di lunghe catene di strutture coordinate legate ad un unico elemento testa, quali ad esempio le liste, in questi casi considerare queste strutture come un unico periodo comporta per il parser l'onere di ricostruire delle relazioni di dipendenza molto lunghe (anche 100 tokens).

4. Citazioni di partizioni interne di testi legislativi (paragrafo, articolo, comma) gerarchicamente organizzate. Sono trattate come catene di modificatori a cascata governate da una testa nominale.

5. Aspetti di segmentazione del testo in periodi (a causa della particolare organizzazione testuale del documento giuridico, ad esempio gli elenchi di disposizioni separate da ";", introdotti da ":").

6. Nel caso della variazione nella realizzazione sintattica di "modificatori" coordinati che ho riscontrato durante il mio lavoro di correzione manuale delle annotazioni automatiche, in realtà l'approccio standard descritto al **paragrafo 3.2** è inapplicabile.
Nel applicarlo si starebbe decidendo di mantenere l'informazione sulla struttura di coordinazione (disgiuntiva o meno), e di perdere quella della relazione sintattica diretta tra il singolo "modificatore" e la testa.Essendo qualitativamente più significativa la relazione sintattica tra i singoli "modificatori" e la testa e anche per mantenere il più possibile la coerenza delle annotazioni sintattiche, e ancora perché si è convenuto che quest'altra soluzione mantiene correttamente invariato il livello di profondità nell'albero di dipendenza, si è scelto di procedere con l'approccio opposto, e, quindi, in questo caso, di preservare la diversità di realizzazione sintattica tra il *comp* "di parità" e il *mod* "attinenti".

"...

83	violazioni	violazione	S	S	num=p|gen=f	78	disj	_	_
84	di	di	E	E	_	83	comp	_	_
85	leggi	legge	S	S	num=p|gen=f	84	prep	_	_
86	di	di	E	E	_	85	comp	_	_
87	parità	parità	S	S	num=n|gen=f	86	prep	_	_
88	o	o	C	CC	_	86	dis	_	_
89	comunque	comunque	B	B	_	90	mod	_	_
90	attinenti	attinente	A	A	num=p|gen=n	85	mod	_	_
91	alla	al	E	EA	num=s|gen=f	90	comp	_	_
92	condizione	condizione	S	S	num=s|gen=f	91	prep	_	_
93	della	di	E	EA	num=s|gen=f	92	comp	_	_
94	donna	donna	S	S	num=s|gen=f	93	prep	_	_

..."

3.4 3.4 Variazione dell'accuratezza del parser nei diversi casi d'analisi

L'approccio all'Adattamento di Dominio nel caso che andremo ad osservare, come già detto è di tipo Data Driven supervisionato.

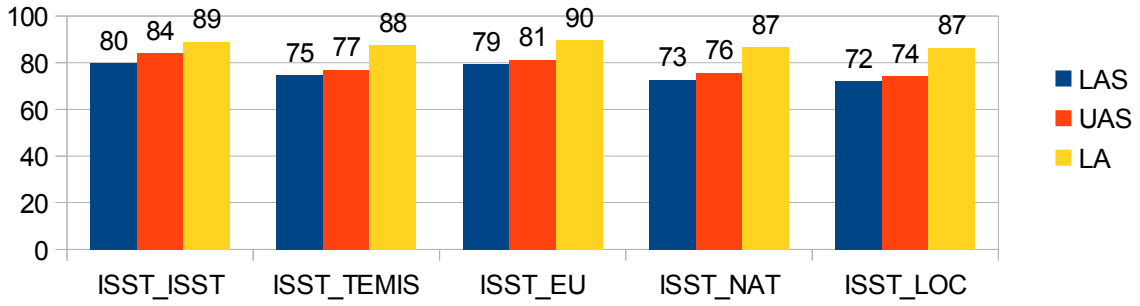
Tabella 3.4

Training Corpus	Target Corpus	LAS	UAS	LA
ISST	ISST	**79,71**	**84,03**	**88,98**
ISST+TEMIS	ISST	**80,62**	**85,03**	**89,14**
TEMIS	ISST	61,78	67,78	77,33
ISST	TEMIS	**74,72**	**77**	**87,61**
ISST+TEMIS	TEMIS	**81,88**	**83,69**	**92,35**
TEMIS	TEMIS	79,32	81,52	90,66
ISST	TEMIS-EU	**79,3**	**81,16**	**89,65**
ISST+TEMIS	TEMIS-EU	**84,27**	**86,18**	**92,55**
TEMIS	TEMIS-EU	80,28	82,87	89,86
ISST	TEMIS-NAT	72,75	75,7	86,66
ISST+TEMIS	TEMIS-NAT	80,01	82,19	91,53
TEMIS	TEMIS-NAT	77,83	80,16	89,85
ISST	TEMIS-LOC	72,19	74,22	86,55
ISST+TEMIS	TEMIS-LOC	81,41	82,73	92,97
TEMIS	TEMIS-LOC	79,88	81,56	92,26

Per valutare (vedi **Tabella 3.4**) l'efficienza dei diversi approcci (Fuori-Dominio, In-Dominio, Combinatio) sono stati utilizzati lo script eval07.pl e il software MaltEval[26] (Jens Nilsson, Joakim Nivre 2008), i quali forniscono informazioni dettagliate sull'accuratezza dell'analisi ai livelli morfologico e sintattico.

3.5 Variazione dell'accuratezza del parser addestrato su ISST

Grafico 3.5 Accuratezze con training ISST



I valori di LAS, UAS e LA nei diversi casi di parsing con training ISST:
target ISST, target TEMIS, target TEMIS-EU, target TEMIS-NAT, e target TEMIS-LOC.

26 http://w3.msi.vxu.se/users/jni/malteval

3.5.1 L'accuratezza e gli errori su ISST

Tabella 3.5.1A

L'accuratezza con training ISST su target ISST				
Parole	Testa Corretta	Dipendenza Corretta	Entrambe Corrette	Entrambe Errate
5165	4340	4596	4117	346
100%	84%	89%	**80%**	7%

L'accuratezza del parser addestrato su ISST e valutato sul target ISST.

Grafico 3.5.1a Accuratezza con training ISST su target ISST

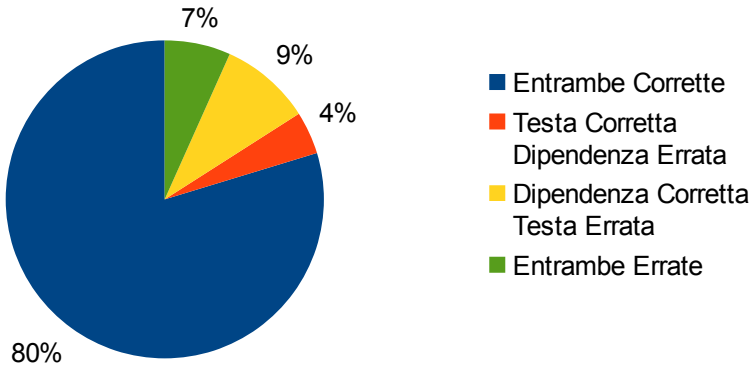
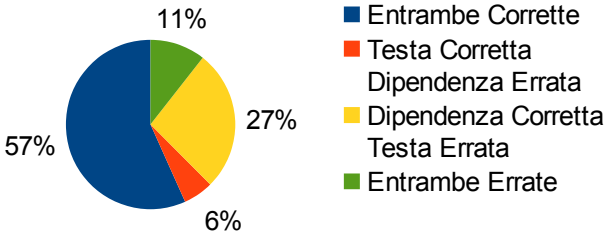
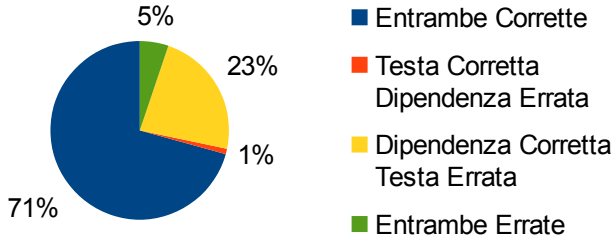


Grafico 3.5.1b
Accuratezza training ISST target ISST
su congiunzioni



Il parser addestrato su ISST, come prevedibile, consegue un'accuratezza elevata su target ISST, in particolare la precision media dell'annotazione del tipo di relazione di dipendenza è 85,3 e la recall media 75,82 mentre i valori di accuratezza LAS, UAS e LA si dimostrano sempre maggiori o uguali all'80%, il che significa che su 10 parole per 8 l'annotazione è completamente corretta, per 1 è corretta solo la relazione di dipendenza e per 1 o è corretta solo la testa, o sono errate entrambe (vedi **Grafico 3.5.1a**). In termini di errore quindi, volendo prescindere dal fatto che siano errate entrambe testa e relazione di dipendenza oppure almeno una delle due sia corretta, su 10 parole 2 soffrono un errore d'annotazione di qualche tipo (20%). È da segnalare un'accuratezza decisamente bassa nell'annotazione di una parte del discorso in particolare, ossia quella delle congiunzioni.

Grafico 3.5.1c
Accuratezza training ISST target ISST su punteggiatura



Solo 6 congiunzioni su 10 sono annotate correttamente e invece nelle altre 4 quasi sempre la testa è errata, in 1 annotazione su 10 oltre alla testa è errata anche la relazione di dipendenza, e in 1 su 20 è errata solo la relazione dipendenza (vedi **Grafico 3.5.1b**). La situazione è simile per la punteggiatura per quel che riguarda l'errore di annotazione della testa (vedi **Grafico 3.5.1c**).

Il quadro emergente è che ci sia un problema di accuratezza relativo principalmente all'annotazione della testa sulle congiunzioni e sulla punteggiatura, infatti, mentre i casi in cui il tipo di dipendenza è errato sono compatibili, o addirittura ridotti (nel caso della punteggiatura) rispetto all'accuratezza generale, la percentuale d'errore sulla testa è il doppio di quella generale.

3.5.2 L'accuratezza e gli errori su TEMIS

Tabella 3.5.2A

L'accuratezza con training ISST su target TEMIS				
Parole	Testa Corretta	Dipendenza Corretta	Entrambe Corrette	Entrambe Errate
5866	4517	5139	4383	593
100%	77%	88%	**75%**	10%

Come prevedibile, per la diversità di dominio, l'accuratezza d'annotazione con training ISST su target TEMIS cala significativamente (vedi **Tabella 3.5.2A**) (-5%). Anche in termini di precision e recall medie dell'annotazione dei tipi di relazione di dipendenza riscontriamo un decremento, la precision scende a 80,08 (-5) la recall a 71,84 (-4).
Su quattro parole 1 è annotata scorrettamente. È sempre sull'annotazione della testa l'errore più frequente, inoltre per questo tipo d'errore il calo è anche il più significativo (-7%, contro un -1% di incremento d'errore per l'annotazione del tipo di dipendenza).
Anche in questo caso l'accuratezza di annotazione sulle congiunzioni è drasticamente inferiore a quella media, è inoltre da segnalare un calo significativo nell'accuratezza di annotazione sulla punteggiatura. Può essere di un certo interesse notare che anche in questo caso la percentuale d'errore sul tipo di dipendenza è coerente con quella riscontrata in generale (12%), mentre la percentuale d'errore sull'annotazione della testa è di nuovo più del doppio di quella generale. In questo caso partiamo già da accuratezze ridotte, quindi la percentuale di annotazioni corrette scende addirittura sotto il 50% (vedi **Grafico 3.5.2b**).

Grafico 3.5.2a Accuratezza con training ISST su target TEMIS

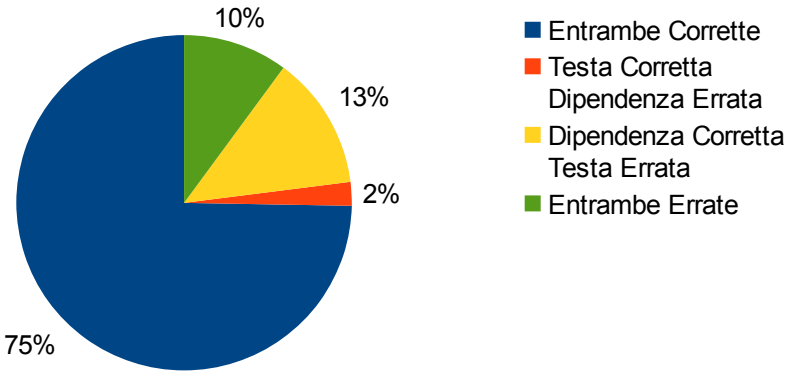


Grafico 3.5.2b
Accuratezza training ISST target TEMIS su congiunzioni

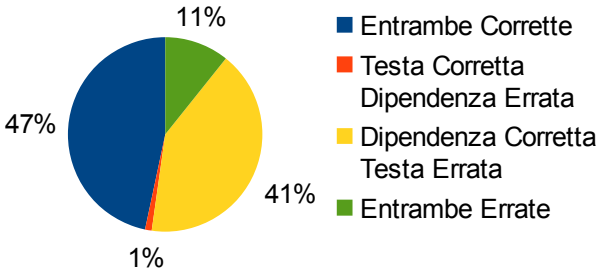
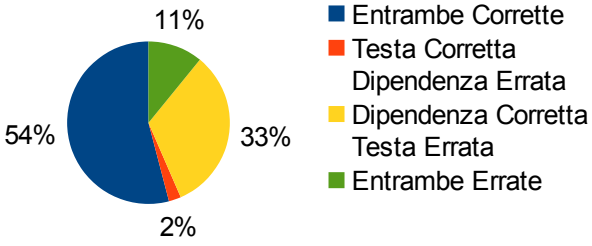


Grafico 3.5.2C
Accuratezza training ISST target TEMIS su punteggiatura



Situazione simile per quanto riguarda l'accuratezza d'annotazione sulla punteggiatura (vedi **Grafico 3.5.2c**): la percentuale d'errore per l'annotazione della testa è praticamente raddoppiata in luogo di una percentuale d'errore sul tipo di dipendenza ancora compatibile con quella generale.
Le percentuali d'errore di questi due casi critici sono di molto superiori a quelle riscontrate per gli stessi casi critici (congiunzioni e punteggiatura) su ISST.
Si passa infatti, per quanto riguarda le percentuali d'errore sull'annotazione della testa, dal 28-38% al 44-53% con un incremento d'errore di circa il 15% (contro un incremento generale del 7%).
Al contrario, nel caso della percentuale d'errore sull'annotazione del tipo di dipendenza sintattica, riscontriamo una variazione quasi nulla, anzi una leggera riduzione. Anche le annotazioni della testa di preposizioni e verbi si dimostrano critiche, con percentuali di errore che passano da 19-17%, con target ISST a 32-30% con target TEMIS, con incremento del 13%

Grafico 3.5.2d
Accuratezza training ISST target TEMIS su preposizioni

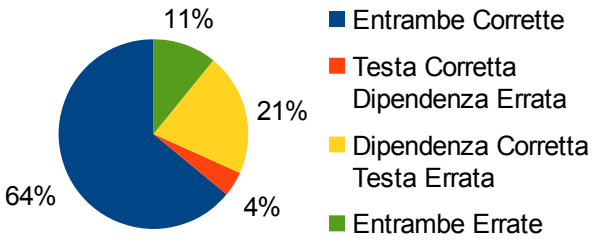
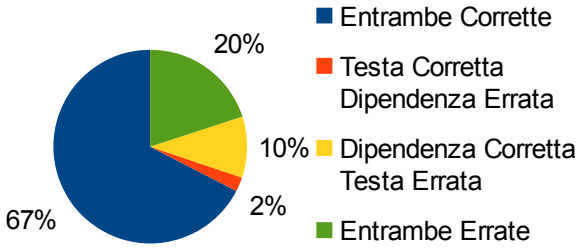


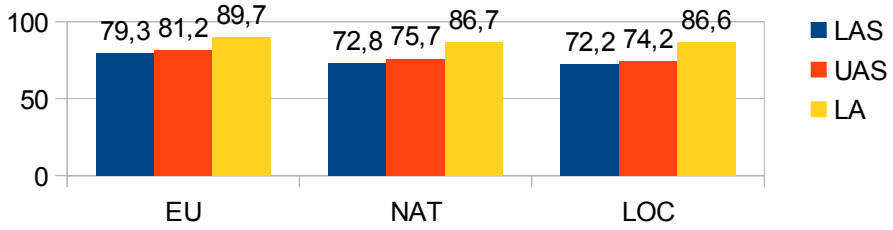
Grafico 3.5.2e
Accuratezza training ISST target TEMIS su verbi



simile a quello riscontrato per congiunzioni e punteggiatura (vedi **Grafico 3.5.2d** e **Grafico 3.5.2e**).
È da rilevare nel caso dei verbi un incremento significativo anche dell'errore di annotazione del tipo di dipendenza (dal 12% al 22%) (vedi **Grafico 3.5.2e**).
Il quadro che emerge è una significativa debolezza sull'annotazione della testa. La generale criticità di analisi di congiunzioni, punteggiatura, preposizioni ed avverbi, dimostra di acuirsi nel contesto del dominio giuridico.

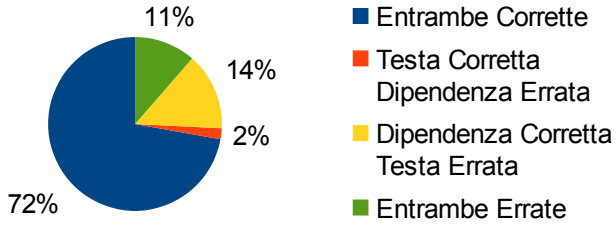
3.5.3 Variazione dell'accuratezza su TEMIS nei sotto-domini giuridici

Grafico 3.5.3 Accuratezza training ISST sui sotto-domini giurid



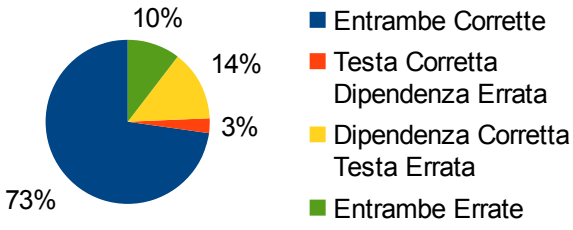
I valori di LAS, UAS e LA nei casi di parsing con training ISST e con i target TEMIS-EU, TEMIS-NAT e TEMIS-LOC

Grafico 3.5.3b
Accuratezza training ISST target TEMIS-LOC



L'accuratezza del parser addestrato su ISST si dimostra nel caso del sotto-dominio europeo (TEMIS-EU), quasi identica all'elevata accuratezza riscontrata su ISST (vedi **Grafico 3.5.3a**). Ciò è interessante in quanto mette in evidenza una considerevole vicinanza di dominio tra ISST-TANL e TEMIS-EU. Possiamo trarne la conclusione che i testi giuridici di legislazione europea parte del corpus TEMIS siano considerevolmente più vicini all'italiano comune giornalistico. Questa constatazione potrebbe dirci molto sul livello di semplicità/chiarezza variabile nei diversi testi giuridici. Da questo riscontro sarebbe forse possibile ricavare ulteriori informazioni per la definizione di basi teoriche per lo sviluppo di strumenti di

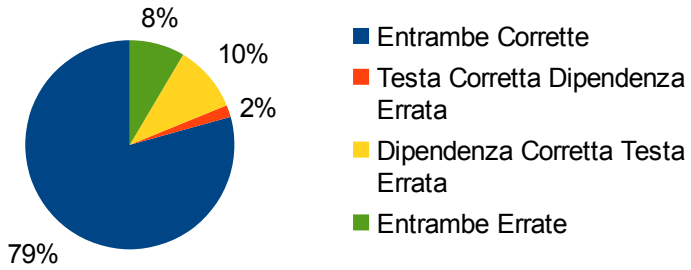
Grafico 3.5.3c
Accuratezza training ISST
target TEMIS-NAT



3.5.3b e **Grafico 3.5.3c**.

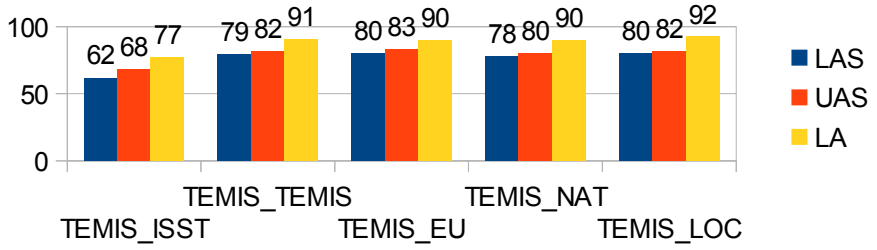
monitoraggio della redazione di atti chiari, semplici e comprensibili basati su indicatori di leggibilità, partendo dal presupposto che se è di più semplice "lettura" per il parser addestrato sulla lingua comune, dovrebbe esserlo anche per i lettori. Negli altri due casi, che si dimostrano molto coerenti tra loro, il decremento di accuratezza è invece addirittura maggiore a quanto riscontrato per TEMIS in generale (vedi **Grafico**

Grafico 3.5.3a
Accuratezza training ISST target TEMIS-EU



3.6 Variazione dell'accuratezza del parser addestrato su TEMIS

Grafico 3.6 Accuratezze con training TEMIS



LAS, UAS e LA con training ISST e con i target ISST, TEMIS, TEMIS-EU, TEMIS-NAT e TEMIS-LOC

3.6.1 L'accuratezza e gli errori su TEMIS

Tabella 3.6.1A

L'accuratezza con training TEMIS su target TEMIS				
Parole	Testa Corretta	Dipendenza Corretta	Entrambe Corrette	Entrambe Errate
5866	4782	5318	4653	419
100%	82%	91%	**79%**	9%

Il parser addestrato su TEMIS consegue, come atteso, dei risultati soddisfacenti sui corpora di dominio giuridico, migliori di quelli che il parser consegue quando addestrato su ISST (del 5% mediamente) (vedi **Grafico 3.6.1a**). Anche la precision media dell'annotazione del tipo di relazione di dipendenza è elevata (89,12) evidenziando le specificità di dominio del training corpus, e la recall media si dimostra, in ogni caso, adeguata, pari a 75,24.
Quest'accuratezza è paragonabile a quelle conseguite dal parser con training ISST sul target ISST e sul target TEMIS-EU.
Congiunzioni, punteggiatura, verbi e preposizioni si confermano quali punti critici, soprattutto per l'annotazione della testa.
Il miglioramento su questi punti è, in ogni caso, significativo ed evidente.

Grafico 3.6.1a
Accuratezza training TEMIS target TEMIS

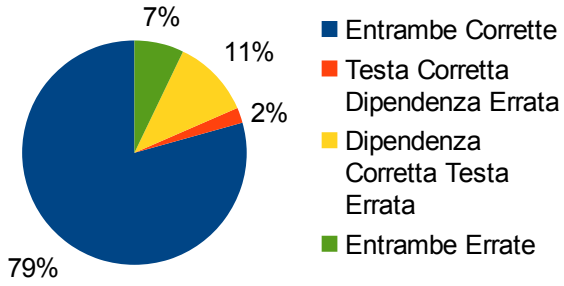
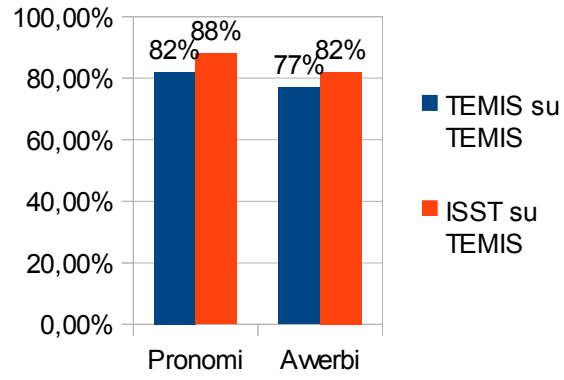


Grafico 3.6.1b
Pronomi e Avverbi su TEMIS
ISST vs TEMIS

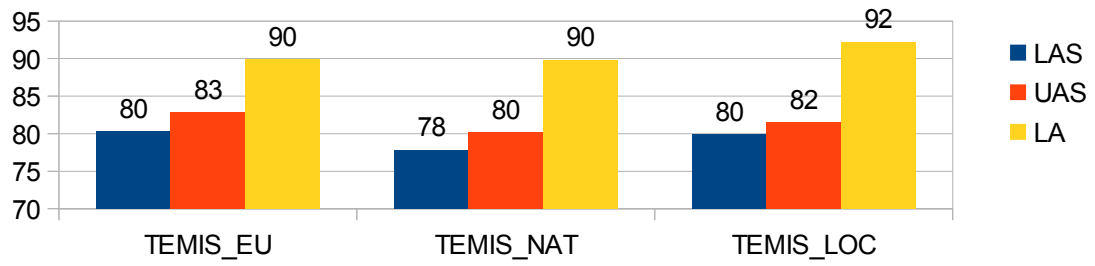


Passiamo da un'accuratezza con training ISST dal 47% al 50% (+3%) per le congiunzioni, dal 54% al 63% (+9%) per la punteggiatura, dal 67% al 70% per i verbi (+3%) e dal 64% al 72% (+8%) per le preposizioni. Un miglioramento d'accuratezza lo abbiamo anche per i sostantivi (+4%) i numerali (+9%). Un miglioramento generale complessivo significativo, purtroppo con effetti collaterali, infatti abbiamo un peggioramento dell'accuratezza per i pronomi (-6%) e per gli avverbi (-5%) (vedi **Grafico 3.6.1b**), nonostante ciò l'accuratezza resta elevata, intorno all'80%. Stabili (e comunque elevate), invece le accuratezze di articoli, aggettivi e determinativi. Il miglioramento ha coinvolto in qualche modo alcuni tratti critici dell'italiano giuridico, le lunghe catene preposizionali, i vizi interpuntivi, e i lunghi periodi ricchi di subordinate e coordinate introdotte da congiunzioni, ma anche la specificità lessicale del linguaggio legale. Il calo in accuratezza (per quanto questa resti elevata) ha invece coinvolto parti del discorso meno ricorrenti per le quali forse una quantità maggiore di frasi annotate (quali ISST può offrire) avrebbe aiutato in termini di copertura, ossia i pronomi (88%→82%), e gli avverbi (82%→77%) (vedi **Grafico 3.6.1b**).

3.6.2 Variazione dell'accuratezza su TEMIS nei sotto-domini giuridici

L'aumento di accuratezza rispetto al parser addestrato su ISST è dell'1% nel caso del sotto-dominio delle leggi europee, del 5% nel caso del sotto-dominio delle leggi statali, del 7% nel caso del sotto-dominio di leggi regionali.

Grafico 3.6.2a Accuratezza training TEMIS su sotto-domini giuridici



I valori di LAS, UAS e LA nei casi di parsing con training TEMIS e con i target TEMIS-EU, TEMIS-NAT e TEMIS-LOC

TEMIS-EU

Si noti lo scarto ridotto (anche in termini di precision: 88 del parser addestrato su TEMIS contro 87,86 di quello addestrato su ISST) nel caso del sotto-dominio delle leggi europee (vedi **Grafico 3.6.2b**), per il quale evidentemente l'addestramento su ISST risulta essere comunque valido.Valido al punto che per quelle parti del discorso generalmente meno ricorrenti (e quindi più a rischio di scarsa rappresentazione in corpora piccoli quali TEMIS) sulle quali l'addestramento sul più vasto ISST si dimostra vincente fornendo un copertura maggiore (in termini di recall infatti l'addestramento su ISST batte quello su TEMIS 82 a 76,4), l'accuratezza rimane nettamente inferiore rispetto a quella del parser con training ISST. Lo scarto è in questo caso addirittura superiore a quello riscontrato per TEMIS in generale (-7% per i pronomi, e -9% per gli avverbi).

Grafico 3.6.2b
Accuratezza training TEMIS
target TEMIS-EU

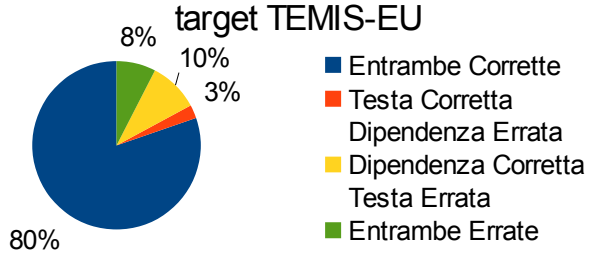
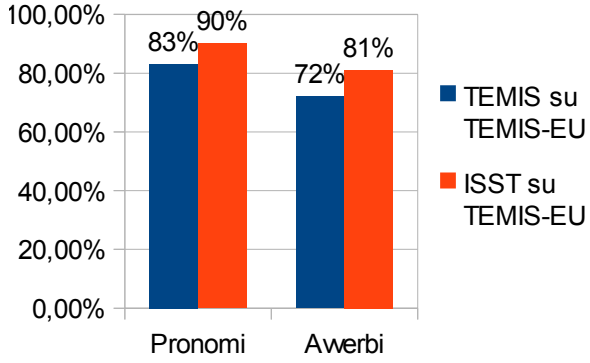


Grafico 3.6.2c
Pronomi e Avverbi in TEMIS-EU:
ISST vs TEMIS



Sostanzialmente i due addestramenti ISST e TEMIS, conseguono, nel caso del sotto-dominio europeo, accuratezze complessivamente simili, nella misura in cui l'addestramento su TEMIS recupera sui tratti specifici di dominio, consentendo di annotare correttamente sostantivi (+1%), preposizioni (+6%), verbi (+1%), (fanno eccezione, stranamente, la punteggiatura e le congiunzioni, ciò potrebbe indicare una ridotta occorrenza di vizi interpuntivi e di lunghe catene di coordinate nel sotto-domino europeo) e l'addestramento su ISST recupera in forza della vastità del corpus, utile a rendere il parser capace di annotare correttamente parti del discorso generalmente meno ricorrenti (vedi **Grafico 3.6.2c**). Nessuno dei due addestramenti si dimostra, in definitiva, più vantaggioso.

TEMIS-LOC

Grafico 3.6.2d
Accuratezza training TEMIS
target TEMIS-LOC

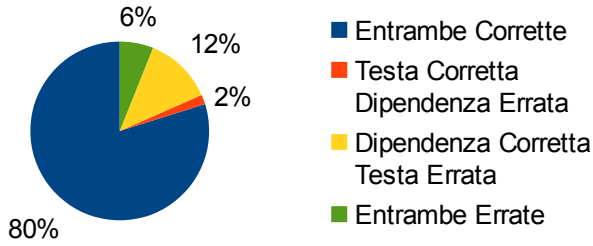
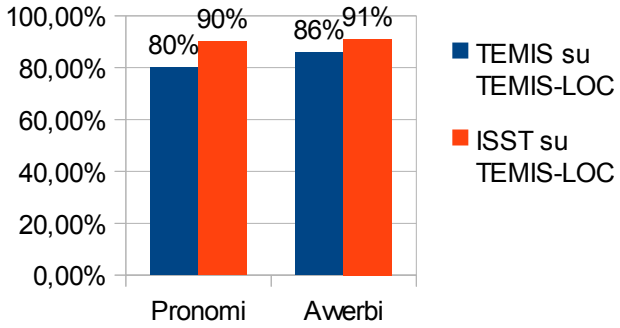


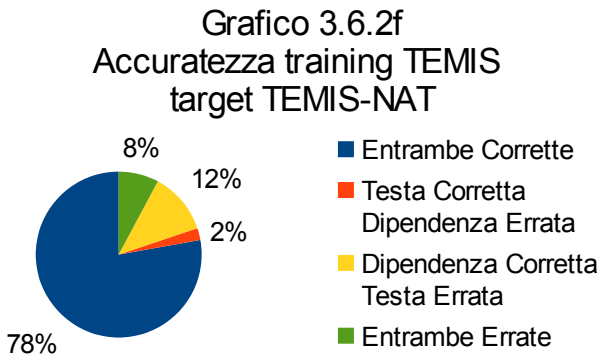
Grafico 3.6.2e
Pronomi e Avverbi su TEMIS-LOC
ISST vs TEMIS

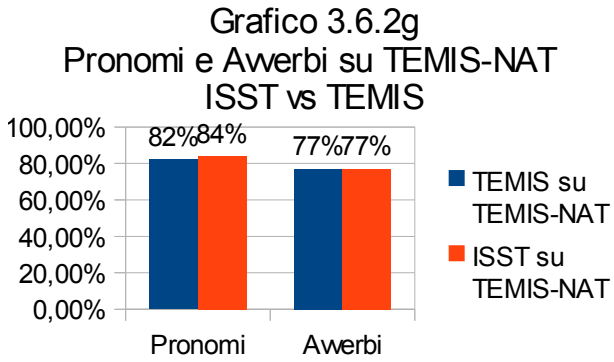


Il sotto-dominio delle leggi regionali (TEMIS-LOC), risulta in qualche modo il più lontano dall'italiano comune.
Per questo sotto-dominio l'incremento in accuratezza con training TEMIS ha i valori maggiori (+8%) (vedi **Grafico 3.6.2d**), mentre l'accuratezza con training ISST, sempre su questo sotto-dominio, è la minore riscontrata (72%) (vedi **Grafico 3.5.3b**).
Il miglioramento d'accuratezza è distribuito principalmente sulle parti del discorso più intaccate dalla variazione di dominio: sostantivi (+8%), preposizioni (+10%), punteggiatura (+17%), verbi (+8%), numerali (+9%), congiunzioni (+7%).
Nonostante la distanza netta dall'italiano comune, esplicita nei tratti sopra citati, rimane decisamente superiore, anche per questo corpus, l'accuratezza del parser addestrato su ISST nell'annotazione di pronomi (-10%) e avverbi (-5%) (vedi **Grafico 3.6.2e**).
Diversamente da quanto riscontrato per TEMIS-EU, in questo caso i tratti specifici di dominio incidono al punto che il miglioramento su questi rende (seppur con effetti collaterali) complessivamente l'addestramento su TEMIS significativamente più vantaggioso di quello su ISST (+8%).

TEMIS-NAT

Per quanto riguarda il sotto-dominio delle leggi statali, riscontriamo alcune differenze rispetto a quanto osservato per TEMIS in generale e per TEMIS-EU e TEMIS-LOC.
Infatti in questo caso, pur avendo gli attesi incrementi di accuratezza su sostantivi (+5%), preposizioni (+9%), punteggiatura (+11%), numerali (+11%) e congiunzioni (+4%), manca l'altrettanto atteso miglioramento nell'accuratezza d'annotazione dei verbi (+0%) e non riscontriamo il calo significativo d'accuratezza su pronomi (solo -2%) e avverbi (-0%) (vedi **Grafico 3.6.2g**).
È nella norma rispetto alla media su TEMIS l'incremento generale dell'accuratezza (+5%) (vedi **Grafico 3.6.2f**).

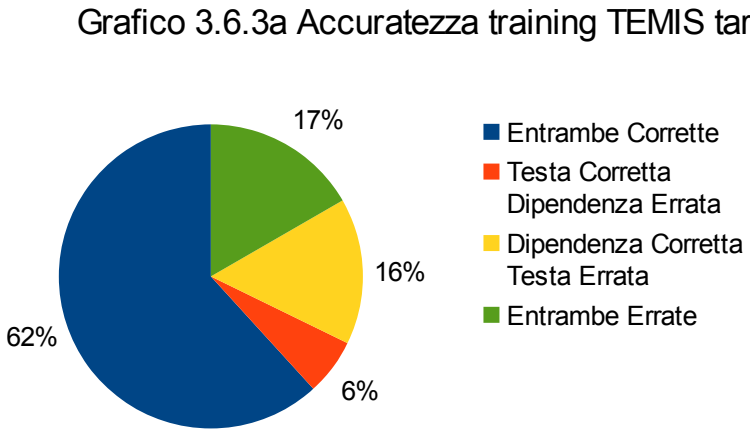




Questa volta l'ago della bilancia pende decisamente in favore dell'addestramento su TEMIS, infatti, in questo caso, se da un lato riscontriamo in misura minore gli incrementi di accuratezza dovuti alle specificità di dominio, dall'altro non riscontriamo affatto gli "effetti collaterali" su pronomi e avverbi (vedi **Grafico 3.6.2g**). Questo è il sotto-dominio per il quale l'addestramento del parser su TEMIS sembra ottenere migliori risultati.

3.6.3 L'accuratezza e gli errori su ISST

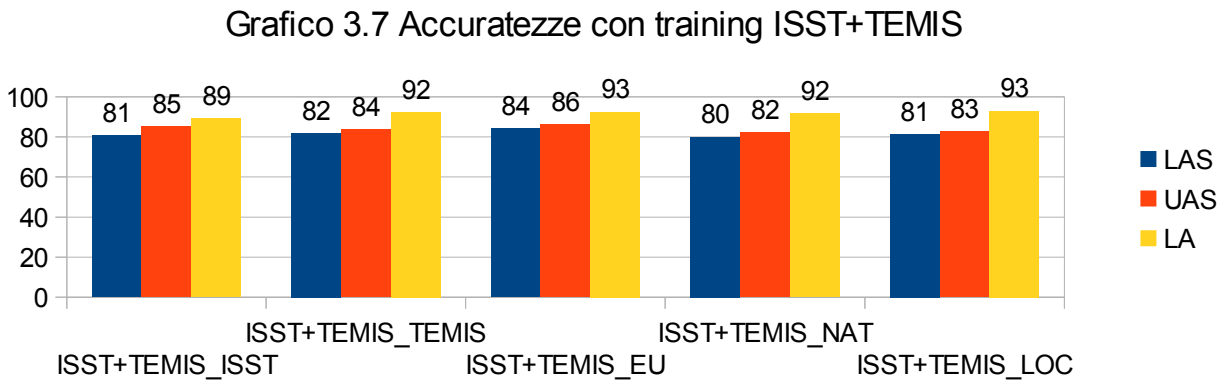
Il parser addestrato su TEMIS consegue dei risultati insoddisfacenti sul corpus dell'italiano giornalistico (una perdita in accuratezza generale del 18%) (vedi Grafico **3.6.3a**). Questo è in effetti un risultato atteso. Valori bassi di precision (74,64) e recall (67,94) medie dell'annotazione del tipo di dipendenza confermano la tendenza osservata.



Il decremento dell'accuratezza è riscontrato per sostantivi (-18%), per la punteggiatura (-26%), per le preposizioni (-16%), per i verbi (-24%), per gli avverbi (-16%), per i pronomi (-27%), per le congiunzioni (-20%), per i numerali (-28%), reggono articoli (-3%), aggettivi (-3%), e determinativi (-3%).
Questo calo vistoso, era atteso, e trova ovvie spiegazioni nell'inadeguatezza del training corpus, sia dal punto di vista della specificità del dominio (training giuridico specifico in luogo di un target di italiano comune) con tutto ciò che ne consegue in termini di accuratezza di annotazione su sostantivi e verbi, sia dal punto di vista della copertura ridotta con quel che ne consegue in termini di accuratezza di annotazione sulle parti del discorso meno ricorrenti (pronomi, avverbi etc.).

La strategia dell'adattamento Data-Driven supervisionato In-Dominio se può sembrare idonea ad adattare il parser ad un nuovo dominio, non lo è affatto per mantenere una retro-compatibilità al dominio o ai domini per i quali esso era addestrato precedentemente.
Adottare questo approccio obbligherebbe all'addestramento di diversi parser specifici di dominio, non necessariamente riuscendo ad ottenere nel contempo il massimo dell'accuratezza possibile.

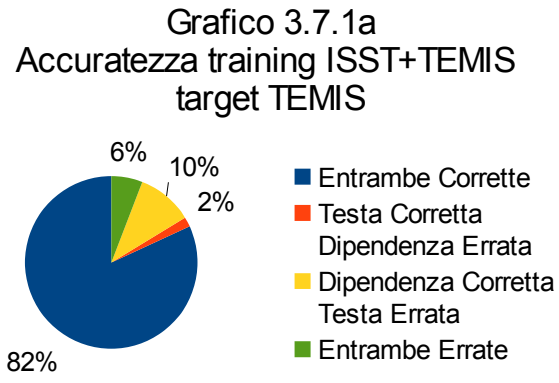
3.7 Variazione dell'accuratezza del parser addestrato su ISST-TANL e TEMIS combinati

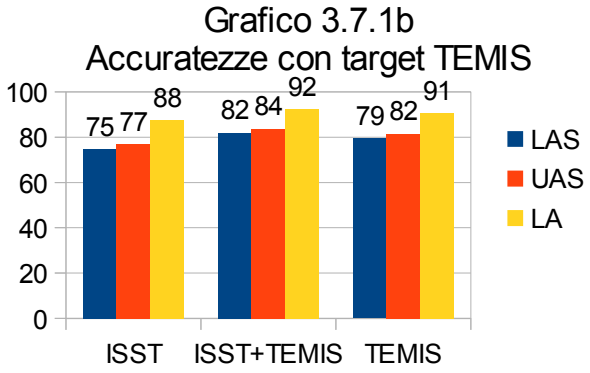


I valori di LAS, UAS e LA nei casi di parsing con training combinato ISST+TEMIS e con i target ISST, TEMIS, TEMIS-EU, TEMIS-NAT e TEMIS-LOC.

Ad un primo riscontro, osserviamo che, estendendo ISST con TEMIS e addestrando il parser su questo corpus combinato, l'accuratezza migliora a prescindere dal target corpus (vedi **Grafico 3.7**).

3.7.1 L'accuratezza e gli errori su TEMIS

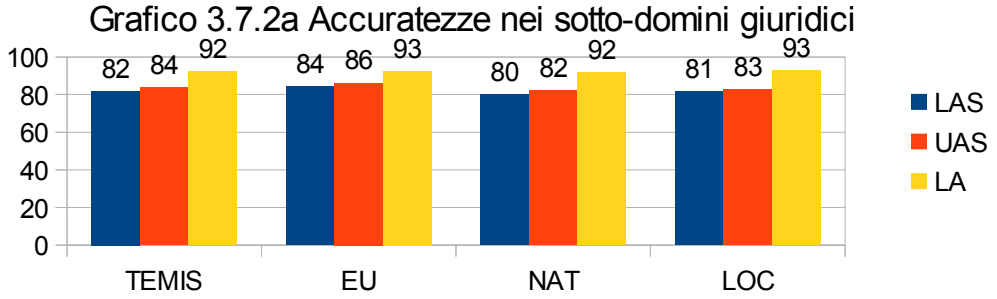




L'accuratezza su TEMIS è quella che ci interessa in particolar modo per misurare l'efficacia dell'approccio all'adattamento al dominio giuridico.
In termini di accuratezza complessiva l'addestramento combinato sembra offrire i risultati migliori (vedi **Grafico 3.7.1a** e **Grafico 3.7.1b**).

La precision media dell'annotazione dei tipi di relazione di dipendenza, non si dimostra in realtà particolarmente elevata 83,35 (inferiore a quella riscontrata con training TEMIS sullo stesso target), mentre la recall risulta più soddisfacente 78,56 (superiore a quella riscontrata con training TEMIS). Analizzando le accuratezze anche per parte del discorso e confrontandole con quelle conseguite dal parser addestrato su TEMIS (che sin qui aveva conseguito le accuratezze migliori su questo target) risulta un miglioramento in tutti i casi a prescindere dalla parte del discorso (tranne, rispetto al parser addestrato su TEMIS, un calo del -1% per la punteggiatura e, rispetto al parser addestrato su ISST, un calo del -2% per gli avverbi e uno del -3% per i pronomi).
Intuitivamente la spiegazione di questo miglioramento (anche nel caso di pronomi e avverbi è un miglioramento rispetto all'accuratezza del parser addestrato su TEMIS) delle accuratezze in tutti i casi, può essere ricondotta per quanto riguarda i tratti più specifici di dominio alla parte di training costituita da TEMIS e, per quanto riguarda la copertura, alla parte di training corpus costituito da ISST.
Come già discusso, risulta evidente che l'apporto di una grande quantità di frasi Fuori-Dominio influisca positivamente in misura significativa.
È significativo che l'annotazione migliori, non solo rispetto all'accuratezza conseguita con l'addestramento dal solo ISST, ma anche rispetto all'accuratezza conseguita con l'addestramento dal solo TEMIS (2% circa nei casi dei sotto-domini delle leggi statali e regionali), sancendo l'approccio con addestramento combinato come più efficiente.

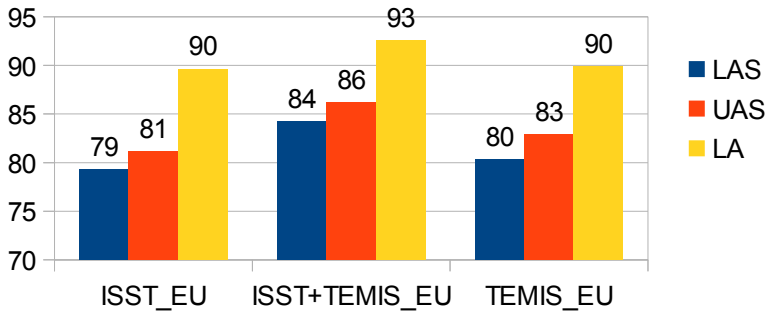
3.7.2 Variazione dell'accuratezza su TEMIS nei sotto-domini giuridici



I valori di LAS, UAS e LA nei casi di parsing con training combinato ISST+TEMIS e con i target TEMIS, TEMIS-EU, TEMIS-NAT e TEMIS-LOC

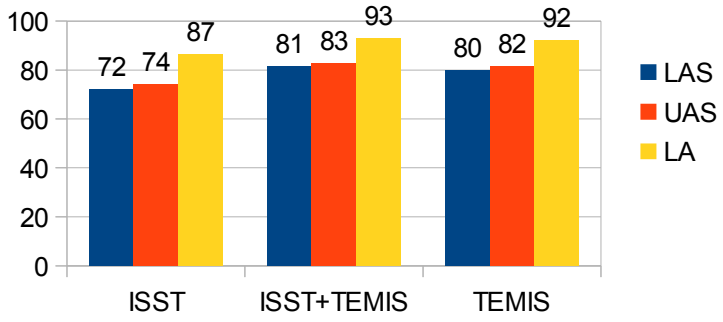
È interessante osservare, ancora, come il miglioramento rispetto all'addestramento con il solo TEMIS sia particolarmente accentuato (+4%) nel caso del sotto-dominio delle leggi europee, pari al doppio e più del miglioramento negli altri due sotto-domini (+1% e +2%) evidenziando come l'aggiunta di ISST apporti un miglioramento in particolare su questo sotto-dominio (vedi **Grafico 3.7.2a**).
Abbiamo già riscontrato in precedenza come il parser addestrato su ISST abbia un'accuratezza abbastanza elevata su TEMIS-EU nonostante la diversità di dominio.
Laddove il sotto-dominio specifico risulti essere più simile linguisticamente al dominio estraneo di cui si dispone una grande quantità di frasi è possibile che il miglioramento sia più accentuato.

Grafico 3.7.2b Incremento su EU



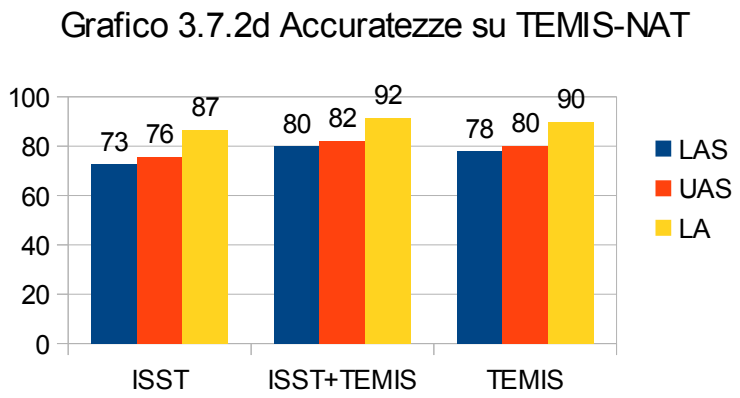
È un risultato molto soddisfacente che laddove il parser con i due addestramenti su ISST e su TEMIS separatamente, conseguiva un'accuratezza di circa l'80%, con l'addestramento sui due corpora combinati, consegua un'accuratezza dell'84% (vedi **Grafico 3.7.2b**). Il miglioramento è riscontrato per tutte le parti del discorso, eccezioni marginali sono, rispetto alle accuratezze conseguite dal parser addestrato su ISST, gli articoli (-1%) e i pronomi (-2%) (comunque in situazione di miglioramento rispetto alle accuratezze del parser addestrato su TEMIS).

Grafico 3.7.2c Accuratezze su TEMIS-LOC



I valori di LAS, UAS e LA nei casi di parsing con target TEMIS-LOC con i diversi training ISST, ISST+TEMIS combinati e TEMIS

Su TEMIS-LOC abbiamo un miglioramento poco significativo ma generalizzato rispetto all'addestramento su TEMIS, con l'eccezione dei numerali (-1%) e dei pronomi (-2% rispetto al parser addestrato su ISST). Il miglioramento medio è dell'1% (vedi **Grafico 3.7.2c**). Questo è il sotto-dominio per il quale l'addestramento combinato ISST+TEMIS incide meno proficuamente in termini di incremento dell'accuratezza.

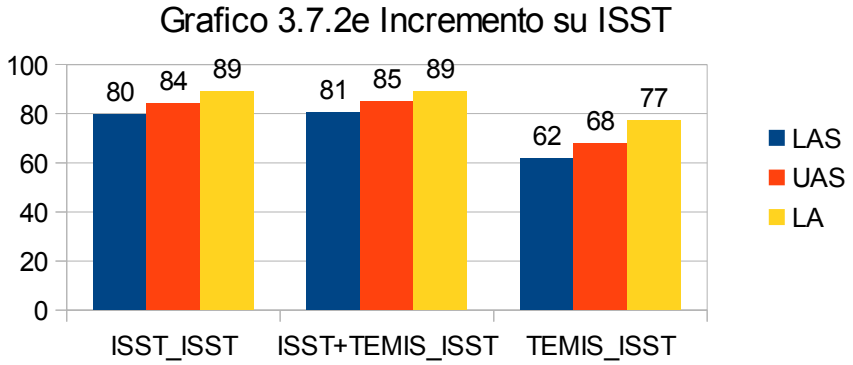


I valori di LAS, UAS e LA nei casi di parsing con target TEMIS-NAT con i diversi training ISST, ISST+TEMIS combinati e TEMIS

Anche su TEMIS-NAT abbiamo un miglioramento generale dell'accuratezza (+2%) (vedi **Grafico 3.7.2d**), anche se in questo caso riscontriamo un debole decremento di accuratezza nel caso dei pronomi (-3% rispetto al parser addestrato su ISST, -1% rispetto al parser addestrato su TEMIS), ed uno più significativo nel caso degli avverbi (-6%). In ogni caso, anche tenuto conto dei limitati cali di accuratezza, l'addestramento combinato ISST+TEMIS si dimostra un approccio efficiente all'adattamento di dominio, risolvendo in maniera complementare sia le esigenze indotte dalle specificità di dominio, sia le esigenze di copertura.

3.7.3 L'accuratezza e gli errori su ISST-TANL

Cosa aspettarci come accuratezza su ISST è meno ovvio e scontato che per i casi analizzati. È interessante scoprire che l'accuratezza del parser addestrato su ISST+TEMIS combinati, sul target ISST, non solo migliora rispetto a quella del parser addestrato sul solo TEMIS, ma addirittura supera quella del parser addestrato sul solo ISST che per questo target consideravamo un punto di riferimento indiscutibile.



I valori di LAS, UAS e LA nei casi di parsing con target ISST con i diversi training ISST, ISST+TEMIS combinati e TEMIS

È stato riscontrato un miglioramento medio dell'accuratezza dell'1% circa (su target ISST) (vedi **Grafico 3.7.2e**), rivelando, forse, che piccole quantità di frasi di dominio specifico possono contribuire a rendere il sistema capace di influire positivamente, magari

risolvendo eventuali particolarità linguistiche di sotto-dominio presenti in corpora generici rappresentativi dell'intera lingua (in questo caso: eventuali catene di coordinate, lunghi periodi ricchi di subordinate etc.). L'unica eccezione è il caso dei pronomi per i quali si riscontra un calo poco significativo del -2%. Anche precision (85,82) e recall (76,13) mostrano un miglioramento della qualità d'analisi con uno scarto di 1 punto sui rispettivi valori nel caso del training ISST. È certo che quantità delle risorse linguistiche e pertinenza di dominio risultano essere parametri fondamentali per una buona accuratezza del parser.
Aldilà dell'incremento di accuratezza, è di grande importanza che con questo approccio si possa preservare (almeno più che con l'addestramento dal solo TEMIS) la retro-compatibilità del parser nei confronti del dominio sul quale era addestrato precedentemente.

3.8 Confronto tra gli approcci In-Dominio, Fuori-Dominio e Combinato

Nel contesto dell'adattamento di dominio Data-Driven supervisionato l'approccio dell'addestramento combinato su ISST+TEMIS si è dimostrato il più efficiente.
Si è riscontrata la capacità di questo approccio di risolvere almeno in buona parte i deficit di accuratezza riscontrati con l'approccio In-Dominio nel caso dell'addestramento del parser su TEMIS (errori nell'annotazione delle teste di dipendenza di pronomi e avverbi), e quasi totalmente quelli riscontrati con l'approccio Fuori-Dominio (errori nell'annotazione delle teste di dipendenza di sostantivi, preposizioni, punteggiatura, congiunzioni etc.).
Si è riscontrato come quest'approccio, oltre a rivelarsi più che adeguato a conseguire la migliore accuratezza possibile nell'annotazione del nuovo dominio, raggiunga prestazioni soddisfacenti anche nell'annotazione del dominio di partenza, senza quindi inficiare la validità del parser nei confronti di domini diversi da quello per il quale si sta procedendo all'adattamento.

4.

Conclusioni
TEMIS nello shared task di SPLeT 2012

Il 4° workshop sul tema "Elaborazione semantica dei testi giuridici" (“Semantic Processing of Legal Texts”, Splet-2012) presenta il primo compito multilingue condiviso sull'Analisi a dipendenza di testi giuridici.
Tuttavia, da Gildea in poi è un fatto ampiamente riconosciuto che i parser di analisi sintattica a dipendenza allo stato dell'arte soffrano di un drammatico calo di precisione se valutati su domini diversi da quello per cui sono stati addestrati e sviluppati.
Per ovviare a questo problema, negli ultimi anni si è visto un crescente interesse nello sviluppo di metodi e tecniche volti ad adattare i sistemi di analisi attuali a nuovi domini.
Diverse iniziative sono state organizzate per affrontare questo problema: ad esempio, il compito di "Adattamento di Dominio", organizzato nel quadro del CoNLL 2007 (Sagae e Tsujii, 2007a), o il Workshop ACL sull'"Adattamento di dominio per il Natural Language Processing "(DANLP, 2010).
In questo contesto, si inserisce il task di “Adattamento di Dominio” (Dell'Orletta et al., 2012) organizzato nel quadro della terza campagna di valutazione del Natural Language Processing, Evalita-2011[27], di cui si è trattato sopra.
Con la sola eccezione di questo task i cui risultati hanno fornito rilevanti feedback in questa direzione, sono stati finora molto pochi i tentativi portati avanti di quantificare le prestazioni dei parser di dipendenza su testi giuridici, probabilmente a causa dell'indisponibilità di corpora “gold” di dominio giuridico annotati manualmente con informazioni sintattiche, da prendere come punto di riferimento per effettuare la valutazione.
Le eccezioni esistono oltre che per l'Italiano solo per il Tedesco.
È il caso del corpus tedesco di 100 frasi estratte da sentenze giudiziarie e annotato sintatticamente a mano, come descritto da Walter (2009).
Tuttavia, questo corpus è attualmente codificato in un formato d'annotazione nativo del parser (Braun, 2003), quindi per procedere ad una qualunque valutazione del parser di dipendenza sarebbe necessaria la conversione in un formato standard (ad esempio CoNLL).
Per la lingua italiana oltre a TEMIS esiste la parte della Treebank dell'Università di Torino (TUT) (sviluppata presso l'Università di Torino, comprende una sezione del Codice Civile e consta 28.048 tokens; 1100 frasi) annotata con informazioni di dipendenza sintattica.
TEMIS (Venturi, 2012), oltre ad essere stato ulteriormente esteso (estensione alla quale ho partecipato annotando semi-automaticamente dal punto di vista morfosintattico e sintattico frasi del corpus di dominio giuridico) già per Evalita 2011 è stato anche parzialmente annotato semanticamente (nel contesto del mio tirocinio ho partecipato anche a questa annotazione, in questo caso manualmente utilizzando il software SALTO[28]) nel modello FrameNet con frame deontici (9,273 token per un totale di 226 frasi annotate semanticamente)[29] per il task condiviso di SpLet 2012.

27 http://evalita.fbk.eu/index.html
28 Erk et al.,2003bl
29 *Giulia Venturi* (2012), Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts, in Proceedings of the 4th Workshop "Semantic Processing of Legal

Nonostante sia TUT che TEMIS siano corpora di dominio giuridico, in realtà essi rappresentano due diverse sub-varietà di dominio.
In particolare, TEMIS si presta maggiormente a rappresentare le peculiarità del dominio giuridico, infatti, secondo uno dei principali studiosi italiani di lingua giuridica Garavelli (2001), gli articoli del Codice Civile sono meno rappresentativi della complessità linguistica del gergo *legalese* rispetto ad altri tipi di testi legislativi quali leggi, decreti, regolamenti, ecc.
Ciò è confermato dai risultati ottenuti nel compito di "Analisi di dipendenza" di Evalita 2011 (Bosco e Mazzei, 2012) in cui tutti i parser partecipanti hanno mostrato, quando testati sul corpus di frasi del Codice Civile, prestazioni migliori rispetto a quelle conseguite quando testato su corpus di frasi dell'Italiano giornalistico.
Come invece risulta evidente, da quanto sopra esposto in questo lavoro oltre che sulla base delle analisi svolte in occasione del task di "Adattamento di Dominio" di Evalita 2011 (Dell'Orletta et al., 2012), per analizzare in maniera affidabile TEMIS e in generale testi giuridici, è necessario che i parser siano adattatati.
In occasione del task condiviso organizzato nell'ambito del 4° workshop Splet-2012 il parsing a dipendenza dei testi di legge mira a:

- fornire definizioni comuni e coerenti di attività e criteri di valutazione al fine di individuare le sfide specifiche poste dall'analisi di questo tipo di testi in diverse lingue,
- delineare un idea più chiara delle prestazioni attuali dei sistemi di analisi allo stato dell'arte,
- condividere risorse multi-linguistiche specifiche di dominio.

Il compito comune è stato organizzato in due diverse attività come descritto di seguito:
1. Parsing a Dipendenza: questo rappresenta l'attività basilare propedeutica, con particolare attenzione all'analisi della dipendenza dei testi giuridici.
Quest'attività è volta a verificare le performance dei sistemi di analisi sui testi normativi;
2. Adattamento Domain: questa è un'attività più impegnativa (e opzionale) con particolare attenzione all'adeguamento del parser generale di dipendenza al dominio giuridico. Quest'attività è volta a indagare i metodi e le tecniche per l'estrazione automatica di informazioni da grandi corpora non annotati del dominio di destinazione per migliorare le prestazioni dei sistemi di analisi generali in materia di testi giuridici.
Le lingue trattate sono l'inglese e l'italiano.
La Valutazione è stata effettuata in termini di misure standard della precisione del parsing a dipendenza, vale a dire LAS rispetto ad un insieme di test-corpora del dominio giuridico.
Il task di Splet 2012 è stato la prima gara di parsing a dipendenza di testi giuridici.
In questo contesto, diversi sistemi di analisi sono stati testati contro set di dati legali italiani e inglesi.
Per le due lingue sono stati ottenuti risultati diversi.
Una significativa diminuzione nella precisione è stata osservata nel caso dei testi in lingua inglese, mentre per l'italiano due su dei tre sistemi partecipanti non hanno mostrato alcun calo di precisione quando testati sul corpus di testi giuridici europei, come riscontrato anche in questo lavoro (sebbene con le dovute precisazioni di compensazione di accuratezze su diverse parti del discorso in base al fatto che l'addestramento fornisse maggiore copertura, o maggiore specificità di dominio), probabilmente a causa di particolari caratteristiche di questa sub-varietà del dominio giuridico che ne fanno una

Texts" (SPLeT 2012), held in conjunction with LREC 2012, Istanbul, Turkey, 27th May, 2012.

varietà più simile all'italiano giornalistico.
Eccezion fatta per questa sub-varietà le prestazioni di tutti i sistemi partecipanti sembrano diminuire in modo significativo (sia per quanto riguarda la sub-varietà di testi legislativi nazionali, sia per quella di testi legislativi regionali).
Tutti i partecipanti hanno utilizzato parser statistici basati su algoritmi di apprendimento automatico, quindi il loro rendimento diminuisce all'aumentare dei tratti linguistici specifici presenti che si verificano poco o per niente nel set di training.
I risultati della valutazione e delle conclusioni finali sono promettenti e incoraggianti per il futuro di applicazioni di estrazione delle informazioni da testi legali.
Lo sviluppo di corpora annotati (anche se pochi ad oggi) del dominio giuridico e le descrizioni dei parser partecipanti, dei loro training e delle loro prestazioni, rappresentano ricche risorse che consentiranno di ottenere un progressivo miglioramento.
È certo, per quanto riscontrato in particolare con l'accuratezza dell'approccio con addestramento combinato su ISST-TANL e TEMIS, sul corpus di dominio giuridico di testi europei, che la specificità di dominio e le grandi quantità di risorse annotate, sono caratteristiche entrambe fortemente incidenti, da un lato per risolvere l'analisi nei casi critici delle specificità di dominio, dall'altro per fornire una copertura la più ampia possibile.
L'approccio ha per altro dimostrato di consentire ottimi margini di retro-compatibilità con il dominio originario di addestramento, limitando il rischio che il progressivo intervento di adattamento possa inficiare l'accuratezza sui domini diversi da quello per il quale si applica la procedura di adattamento di dominio.
Forse la combinazione dei corpora d'addestramento, porterà a risultati migliori, all'aumentare della dimensione del corpus giuridico, comprendendo in sé gli incrementi necessari in specificità di dominio e in copertura allo stesso tempo.
Lavorare ancora sull'estensione delle risorse linguistiche annotate manualmente per il dominio giuridico, come si è già cominciato a fare, per quanto possa essere un lavoro lungo e poco economico, non può che essere una valida linea d'azione che colga l'importanza di questi due diversi elementi, complementari nel conseguire un'accuratezza d'analisi soddisfacente.

Bibliografia

Montemagni et al., 2003
S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi e R. Delmonte. Building and using parsed corpora. In A. Abeill´e, editore, Building and using Parsed Corpora, Language and Speech Series, pp. 189–210. Kluwer, Dordrecht, 2003.

Lenci, Montemagni e Venturi, 2012
Alessandro Lenci, Simonetta Montemagni, Giulia Venturi, Maria Rosaria Cutrullà, "Enriching the ISST–TANL Corpus with Semantic Frames", 2012

Dell'Orletta, 2009
F. Dell'Orletta. Ensemble system for part–of–speech tagging. In Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian), Reggio Emilia, 2009.

Attardi et al.,2009
G. Attardi, F. Dell'Orletta, M. Simi e J. Turian. Accurate dependency parsing with a stacked multilayer perceptron. In Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian), Reggio Emilia, 2009.

Venturi, 2012
Giulia Venturi, Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts, in Proceedings of the 4th Workshop "Semantic Processing of Legal Texts" (SPLeT 2012), held in conjunction with LREC 2012, Istanbul, Turkey, 27th May, 2012.

Biber e Conrad, 2009
D. Biber and S. Conrad. 2009. Register, genre, and style. Cambridge, Cambridge University Press.

Erk et al.,2003bl
K. Erk, A. Kowalski, and S. Pad`o. 2003. The salsa annotation tool-demo description. In Proceedings of the 6th Lorraine–Saarland Workshop, pages 111–113, Nancy, France.

McClosky et al., 2010
D. McClosky, E. Charniak e M. Johnson. Automatic domain adaptation for parsing. Proceedings of the HLT-NAACL'2010, pp. 28–36, Los Angeles, California, 2010.

McClosky et al. 2006
David McClosky, Eugene Charniak, and Mark Johnson. Effective Self-Training for Parsing. In Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 152–159, Brooklyn, New York. Association for Computational Linguistics. 2006

Plank e Van Noord 2011
B. Plank e G. van Noord. Effective measures of domain similarity for parsing. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, pp. 1566–1576, Portland, Oregon, 2011.

Sekine 1997
Satoshi Sekine. The Domain Dependence of Parsing. In In Proceedings of the Fifth Conference on Applied Natural Language Processing, pages 96–102, Washington D.C. 1997

Ratnaparkhi 1999
Adwait Ratnaparkhi. Learning to parse natural language with maximum entropy models. *Machine Learning*, 34(1-3):151–175. 1999

Gildea 2001
D. Gildea. Corpus variation and parser performance. Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001), pp. 167–202, Pittsburgh, PA, 2001.

Clegg e Shepherd 2005
A. B. Clegg e A. J. Shepherd. Evaluating and integrating treebank parsers on a biomedical corpus. Proceedings of the Workshop on Software, pp.14–33, Ann Arbor, Michigan, 2005.

Hara et al. 2005
Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. Adapting a Probabilistic Disambiguation Model of an HPSG Parser to a New Domain. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, Natural Language Processing IJCNLP 2005, volume 3651 of Lecture Notes in Computer Science, pages 199–210. Springer Berlin / Heidelberg 2005.

Daumé III, 2007
Hal Daum´e III. Frustratingly Easy Domain Adaptation. In Proceedings of the 45th Meeting of the Association for Computational Linguistics, Prague, Czech Republic 2007.

Blitzer, McDonald e Pereira 2006
John Blitzer, Ryan McDonald, and Fernando Pereira. Domain Adaptation with Structural Correspondence Learning. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia 2006.

Daumé III, Kumar e Saha 2010
Hal Daumé III, Abhishek Kumar, Avishek Saha: Co-regularization Based Semi-supervised Domain Adaptation, 2010

Chang, Connor e Roth 2010
M. Chang, M. Connor and D. Roth. The Necessity of Combining Adaptation Methods. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2010

Asch e Daelemans 2010
Vincent Van Asch and Walter Daelemans, Using Domain Similarity for Perfomance Estimation, 2010

Lippincott et al. 2010
Tom Lippincott, Diarmuid ´O S´eaghdha, Lin Sun, and Anna Korhonen, Exploring variation across biomedical subdomains. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 689–697, Beijing, China, August, 2010

Dell'Orletta e Montemagni, 2010a
F. Dell'Orletta e S. Montemagni. Tecnologie linguistico–computazionali per la valutazione delle competenze linguistiche in ambito scolastico. In *Atti del XLIV Congresso Internazionale di Studi della Societ`a di Linguistica Italiana (SLI 2010)*, 27-29 settembre, Viterbo, 2010a.

Venturi 2011
Giulia Venturi, Lingua e diritto: una prospettiva linguistico-computazionale, Torino 2011

McDonald e Nivre 2007
R. McDonald e J. Nivre. Characterizing the errors of data–driven dependency parsing models. In *Proceedings of the the EMNLP-CoNLL*, pp. 122–131, 2007.

Nilsson e Nivre 2008
J. Nilsson e J. Nivre. Malteval: an evaluation and visualization tool for dependency parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 161–166, Marrakech, Morocco, 2008.

Sagae e Tsujii 2007a
K. Sagae and J. Tsujii. 2007a. Dependency parsing and domain adaptation with lr models and parser ensembles. In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, pages 1044–1050, Prague, 2007

Sagae e Tsujii 2007b
Kenji Sagae and Junichi Tsujii. 2007b. Dependency parsing and domain adaptation with lr models and parser ensemble. In Proceedings of the EMNLP–CoNLL 2007, pages 1044–1050, 2007

Dell'Orletta et al. 2012
F. Dell'Orletta, S. Marchi, S. Montemagni, G. Venturi, T. Agnoloni, and E. Francesconi. Domain adaptation for dependency parsing at evalita 2011. In Working Notes of EVALITA 2011, Rome, Italy, 2012

Braun 2003
C. Braun. Parsing german text for syntactico–semantic structures. In *Prospects and Advances in the Syntax/Semantics Interface, Proceedings of the Lorraine–Saarland Workshop*, Nancy, France, 2003.

Bosco e Mazzei 2012
C. Bosco and A. Mazzei. The evalita 2011 parsing task: the dependency track. In Working Notes of EVALITA 2011, Rome, Italy. 2012