UNIVERSITÀ DEGLI STUDI DI TORINO

Scuola di Dottorato in Studi Umanistici

Dottorato in Studi Euro–Asiatici: Indologia, Linguistica, Onomastica

Indirizzo: Linguistica, Linguistica Applicata e Ingegneria Linguistica (ciclo XXIII)

Lingua e diritto: una prospettiva linguistico–computazionale

Tutor:

prof. Mario SQUARTINI

Candidata:

Giulia VENTURI

Anni Accademici 2008–2011

Indice

Ringraziamenti

Questo studio è il risultato del lavoro da me svolto durante i tre anni di dottorato presso l'Università di Torino. Desidero pertanto ringraziare prima di tutto Mario Squartini e Carla Marello che mi hanno dato l'opportunità di portare avanti le mie ricerche, seguendomi e consigliandomi costantemente lungo tutti questi tre anni.

Nello svolgimento di questo lavoro, l'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR di Pisa ha avuto un ruolo centrale. La mia riconoscenza va dunque a tutte le persone che mi hanno accolto e offerto il supporto indispensabile durante tutto il periodo del dottorato.

Un riconoscimento speciale va alle persone del gruppo di ricerca di cui faccio parte presso l'Istituto di Linguistica Computazionale. In particolare, a Simonetta Montemagni per avermi guidata, consigliata e incoraggiata in ogni momento del mio lavoro e a Felice Dell'Orletta per gli originali spunti di ricerca che mi ha offerto e per il supporto che mi ha dato durante le diverse fasi del mio studio.

La mia sentita riconoscenza va a Simone Marchi e Emiliano Giovannetti, che fin dai primi momenti mi hanno affettuosamente accolta all'Istituto di Linguistica Computazionale.

La catena di strumenti di Trattamento Automatico del Linguaggio usati in questo lavoro è stata sviluppata nell'ambito delle attività congiunte del Dylan Lab (Laboratorio per lo studio delle dinamiche linguistico–cognitive) dell'Istituto di Linguistica Computazionale e dell'Università di Pisa. Un ringraziamento va pertanto a tutte le persone che hanno contribuito alla realizzazione degli strumenti di analisi.

La metodologia di annotazione semantica fondata sulla specializzazione di FrameNet è parte delle attività tutt'ora in corso svolte in modo congiunto dall'Istituto di Linguistica Computazionale del CNR di Pisa e dall'Università

di Pisa nell'ambito del progetto nazionale IFrame[1] finalizzato allo sviluppo di un FrameNet per la lingua italiana. Il mio riconoscimento va in particolare a Alessandro Lenci, per aver creato le condizioni ottimali affinché io entrassi in contatto con l'International Computer Science Institute (ICSI) di Berkeley e con Charles Fillmore, e a Eva Maria Vecchi per aver condiviso con me le sue competenze in materia di annotazione e i suoi dubbi, dando inizio ad un fruttuoso dialogo tra di noi.

Un sincero ringraziamento va a Tommaso Agnoloni, Enrico Francesconi, Maria Teresa Sagri, Pierluigi Spinosa, Daniela Tiscornia dell'Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG) del CNR di Firenze con i quali ho collaborato lungo questi tre anni. A loro sono in particolare riconoscente per avermi guidata nella parte di questo studio dedicata ad aspetti di rappresentazione formale della conoscenza giuridica, nonché in quella nella quale vengono trattate questioni di estensione e specializzazione di FrameNet per il dominio giuridico.

Parte del lavoro di questa tesi è stato svolto nell'ambito di una serie di progetti europei e nazionali che mi hanno dato l'opportunità di elaborare alcune delle idee e delle ricerche qui contenute. La mia gratitudine va pertanto alle persone coinvolte nei progetti europei BOOTSTREP (*Bo*otstrapping of *O*ntologies and *T*erminologies *St*rategic *R*esearch *P*roject, FP6 n. 028099) e DALOS (*Drafting Legislation with Ontology–based Support*, eParticipation project n. 2006/01/024) e nel progetto nazionale "IC.P10 Migrazioni".

Per avermi semplicemente accompagnata in tutto questo il mio grazie va a Roberto.

[1] http://sag.art.uniroma2.it/iframe/doku.php

Capitolo 1

Introduzione

"Il linguista che si occupa di testi giuridici si trova di fronte ad argomenti che sono, o sono stati, materia di discussione da parte degli specialisti del diritto; a questioni di forma linguistica che possono avere importanza essenziale per la soluzione di problemi giuridici"[1]. Con questa riflessione Bice Mortara Garavelli (2001) inizia lo studio che raccoglie le sue ricerche sulle caratteristiche grammaticali e retoriche rintracciabili in testi giuridici italiani.

Si è scelto di introdurre lo studio qui presentato ricordando le sue parole, dal momento che esse ben chiariscono il perché del carattere intrinsecamente interdisciplinare delle analisi condotte, finalizzate a mettere in luce le potenzialità e i limiti di un approccio all'analisi di testi giuridici realizzata con metodi e strumenti linguistico–computazionali.

Questa prospettiva di analisi, da una parte, condivide con i principali studi sulla lingua del diritto l'idea che il linguista interessato allo studio delle caratteristiche della lingua del diritto debba necessariamente "porsi questioni linguistiche in stretta connessione con questioni giuridiche"[2]. Dall'altra, abbraccia la concezione per cui affrontare le "questioni di forma linguistica" sia indispensabile per risolvere con successo le questioni di contenuto e che anzi, come fanno notare i giuristi stessi, "i problemi di significato degli enunciati giuridici possono essere affrontati solo risolvendone i problemi sintattici"[3].

Una tale attenzione per la lingua del diritto è infatti al centro non solo degli studi dei linguisti, ma anche delle attività dei giuristi le quali, secondo Uberto Scarpelli (1969), devono costantemente consistere in operazioni

[1] Garavelli (2001, p. 4).
[2] Garavelli (2001, p. 34).
[3] Jori e Pintore (1995, p. 209).

che "riguardano il linguaggio ed hanno come strumento il linguaggio", dal momento che il giurista "deve determinare e foggiare significati, riconoscere, costruire o ricostruire relazioni semantiche, e sintattiche e pragmatiche". Perché "se c'è un'attività che richieda una consapevolezza linguistica, questa è l'attività dei giuristi".

In questo senso dunque la risoluzione di **questioni linguistiche** è vista come la chiave di accesso per affrontare **questioni semantiche** connesse con l'interpretazione del discorso giuridico. La centralità di un tale punto di vista è chiaramente sostenuta da Norberto Bobbio (1976, p. 306) che così si interroga: "Che altro è [...] l'interpretazione della legge se non l'analisi del linguaggio del legislatore, cioè di quel linguaggio in cui vengono espresse le regole giuridiche?".

Anche nell'ambito delle ricerche in materia di metodi e strumenti dell'intelligenza artificiale applicata al diritto (ambito noto come 'Artificial Intelligence and Law') e, in particolare, nell'ambito delle più recenti attività basate sull'utilizzo di metodi e tecniche di Trattamento Automatico del Linguaggio, è riconosciuta la centralità di un approccio stratificato all'analisi di testi giuridici. La questione è chiaramente messa in luce da McCarty (2009) che, durante il suo intervento al "Worskhop on Natural Language Engineering of Legal Argumentation" (NaLEA2009), riflettendo sull'importanza di annotare linguisticamente testi di legge, rendendone esplicita in modo automatico la struttura sintattica, porta l'attenzione del pubblico sulle possibili applicazioni che possono trarre vantaggio dall'annotazione linguistica[4]. Esse riguardano principalmente la possibilità di usare l'informazione linguistica acquisita come punto di partenza per rappresentare in modo formale il contenuto informativo dei testi, allo scopo di realizzare compiti di estrazione, organizzazione e gestione automatica della conoscenza generali, quali 'Information Extraction', 'Information Retrieval', 'Text Mining', ecc..., e compiti specializzati per il dominio giuridico, quali 'Legal Ontology Learning', 'Legal Reasoning', 'Legal Argumentation Mining', ecc...

Trovare dunque una metodologia di analisi che permettesse di rendere esplicite le relazioni tra la struttura sintattico–grammaticale di un testo giuridico e il modo in cui vi è organizzato il contenuto semantico–informativo è il proposito che ha guidato l'intero studio. Sulla scia delle considerazioni

[4]L'interrogativo posto da McCarty (2009) in quell'occasione era: "Why parse statutes? To extract their logical structure, to refine the semantics of the domain, to develop a domain ontology".

condotte in ambiti di ricerca diversi tra loro, l'aspetto innovativo dello studio qui presentato consiste pertanto nell'affrontare il tema del rapporto tra analisi linguistica di testi giuridici e accesso al loro contenuto informativo

– facendo affidamento sui più recenti e accurati strumenti di Trattamento Automatico del Linguaggio oggi esistenti che, annotando linguisticamente un testo su più livelli di analisi, consentono di rendere esplicita in modo automatico l'informazione linguistica in esso contenuta;

– adottando una metodologia di accesso incrementale al contenuto del testo la quale, a partire dal riconoscimento della terminologia rilevante in esso contenuta, permette di rendere espliciti i principali elementi informativi presenti e il modo in cui essi interagiscono tra loro grazie ad un processo di annotazione semantica del testo condotta sulla base di un modello di rappresentazione e organizzazione formale del significato semantico–lessicale.

Filo rosso conduttore dell'intero studio è il costante approccio incrementale all'indagine dei modi e delle forme del rapporto tra lingua e diritto, articolato in più fasi. Esso permette, da un lato, di mettere in luce il valore autonomo che hanno le singole fasi e, dall'altro, come il loro interagire restituisca uno sguardo completo sul tema. In quest'ottica, le parti nelle quali è organizzato lo studio, prendendo le mosse da comunità di ricerca diverse tra loro, possono anche essere lette in autonomia. È questo infatti il motivo per cui per ognuna di esse sono di volta in volta tracciate alcune considerazioni conclusive.

È la loro successione tuttavia a chiarire il modo in cui si dipana l'argomentazione. Essa mira a dimostrare come lo studio condotto si proponga di suggerire le risposte ad alcuni interrogativi dai quali si è partiti con l'intento di affrontare i seguenti aspetti di indagine:

- quali sono gli aspetti dei rapporti tra analisi linguistica di testi giuridici e analisi del contenuto oggetto, sino ad oggi, di continue e aperte discussioni? La questione è affrontata grazie ad una rassegna ragionata *i)* dei principali temi di dibattito al centro degli studi di linguisti, giuristi e ricercatori in Intelligenza Artificiale e Diritto e *ii)* delle metodologie di analisi più seguite nell'ambito delle diverse comunità di ricerca;

- considerati i ben noti caratteri di difficoltà della lingua del diritto, in che misura gli strumenti di annotazione linguistica automatica del testo sviluppati per analizzare la lingua comune sono accurati nell'analisi

linguistica di testi giuridici? Ponendo al centro della discussione le strutture linguistiche di testi legislativi per le quali gli strumenti sviluppati per il Trattamento Automatico del Linguaggio comune usati in questo studio generano sistematicamente analisi non corrette, è affrontato il tema delle difficoltà connesse con un compito di trattamento automatico della lingua del diritto;

- nonostante tali difficoltà, gli strumenti di annotazione linguistica automatica sono in grado di ricostruire un articolato profilo linguistico di testi giuridici affidabile a tal punto da fornire una conferma quantitativa degli studi condotti in modo manuale dai linguisti? Il modo in cui a partire da una serie di tratti linguistici presenti nel corpus di testi giuridici qui raccolto e rintracciati sulla base dei vari livelli di annotazione linguistica automatica sia possibile individuare alcune delle loro principali caratteristiche morfosintattiche, sintattiche e lessicali è l'aspetto di indagine connesso con un tale quesito e affrontato nell'ambito di questo studio;

- in un'ottica di analisi incrementale del contenuto semantico–informativo dei testi giuridici, in che modo i vari livelli di annotazione linguistica automatica del testo costituiscono il punto di partenza per accedere alla semantica di un testo? La definizione di una metodologia di indagine articolata in una serie di passaggi progressivi, fondamentali per rendere espliciti elementi sempre più complessi del contenuto, è il tema affrontato nella parte di questo studio dedicata a trattare i principali aspetti semantici connesi l'analisi di testi giuridici;

- tenendo conto di quanto affermato da Giovanni Rovere (2005) circa il fatto che è necessario disporre di "modelli di rappresentazione, atti a dar conto di tutti i fatti linguistici presenti nel corpus", qual è il modello di organizzazione e rappresentazione del significato oggi esistente che meglio consente di rendere esplicito il modo in cui i principali elementi linguistico–informativi sono tra loro organizzati nel testo giuridico? La scelta di un modello di rappresentazione contestuale del significato comune e la sua specializzazione come modello di annotazione semantica di testi giuridici è al centro delle indagini finalizzate ad adottare FrameNet come modello sia di descrizione del significato semantico–lessicale sia di rappresentazione della conoscenza contenuta in testi giuridici;

- ed infine, in che modo è possibile verificare come FrameNet sia concretamente utilizzabile in un compito di annotazione semantica di testi giuridici? Il tema è oggetto del caso di studio qui condotto volto ad individuare le principali specializzazioni necessarie per adottare un tale modello, originariamente sviluppato per organizzare e rappresentare il significato contenuto in corpora rappresentativi della lingua comune, per rendere espliciti gli 'obblighi', 'permessi' e 'divieti' presenti in corpus di atti normativi emanati dallo stato italiano qui assunti come rappresentativi della lingua del diritto.

In quanto segue è dunque brevemente riassunto il modo in cui le discussioni relative agli aspetti di indagine ora delineati sono organizzate nelle tre parti in cui si articola questo lavoro.

Parte I

La prima parte è dedicata a presentare quegli aspetti teorici e metodologici relativi allo studio dei complessi rapporti tra lingua e diritto che, essendo oggetto d'interesse di diverse comunità di ricerca, sono al centro dei più accesi dibattiti. Il Capitolo 2 si prefigge di fondare il carattere interdisciplinare di questo studio sul fatto che "di fronte allo stesso oggetto di studio le pertinenze dei due campi, linguistico e giuridico, si intrecciano e si sovrappongono"[5]. A questo scopo, prende le mosse dalle svariate prospettive di ricerca delineate da Bice Mortara Garavelli (2001), aggiornando la sua rassegna con le più recenti attività di ricerca in materia di metodi e strumenti dell'intelligenza artificiale applicata al diritto (ambito noto come 'Artificial Intelligence and Law') e, in particolare, con la descrizione delle attività basate sull'utilizzo di metodi e tecniche di Trattamento Automatico del Linguaggio finalizzati all'annotazione semantica di testi giuridici.

Sono così riportate (nel Paragrafo 2.1), prima di tutto, le discussioni dei linguisti interessati a trovare una metodologia in grado di definire in modo più chiaro di quanto ora non sia i labili confini tra lingua del diritto, lingua comune e linguaggi specialistici oggetto del discorso giuridico, focalizzandosi sia sul livello lessicale di analisi sia su quello morfosintattico e sintattico. Particolare attenzione è dedicata alle più recenti metodologie di analisi basate sull'uso di corpora testuali e su di un approccio comparativo alle studio delle differenze tra collezioni testuali rappresentative delle diverse varietà di lingua.

[5] Garavelli (2001, p. 4).

In un secondo momento, sono descritte (nel Paragrafo 2.2) le attività dei giuristi e dei filosofi del diritto appartenenti alla scuola analitica italiana di filosofia del diritto, *i)* hanno messo in luce il carattere intrinsecamente linguistico del compito principale per un giurista, quello cioè di interpretare la legge che consiste in quest'ottica in una vera e propria analisi del linguaggio nella quale essa è scritta, e *ii)* hanno studiato la stretta relazione tra lingua del diritto, lingua comune e linguaggi specialistici, soprattutto del punto di vista lessicale, proponendo una possibile classificazione dei principali termini che ricorrono in testi giuridici basata sul loro contesto d'uso.

L'attenzione è infine focalizzata sulle ricerche condotte nell'ambito degli studi informatico–giuridici, con l'obiettivo di mettere in particolare evidenza, da un lato (nel Paragrafo 2.3.1), le attività legate alla 'legimatica' sia *i)* come attività volta a sviluppare strumenti di ausilio alla fase di redazione del testo giuridico (legislativo) e di controllo della qualità del testo redatto, sia *ii)* come attività che, utilizzando strumenti di Trattamento Automatico del Linguaggio, è finalizzata a rendere accessibile il testo legislativo da parte di agenti informatici, arricchendolo con metadati informativi relativi sia all'articolato sia al disposto sulla base di modelli strutturali di testi normativi.

Dall'altro, nel Paragrafo 2.3.2, sono descritte le attività in materia di intelligenza artificiale applicata al diritto, mettendo particolarmente in rilievo *i)* come fin dai suoi esordi l'obiettivo di questa disciplina fosse quello di formalizzare strutture concettuali giuridiche basandosi su metodi di Trattamento Automatico del Linguaggio, sebbene gli strumenti allora a disposizione non lo consentissero, e *ii)* come invece negli ultimi anni si stiano diffondendo sempre di più attività basate sull'uso di tali strumenti applicati alla realizzazione di diversi compiti di gestione del contenuto semantico–informativo di corpora di testi giuridici. Particolare attenzione in questo senso è posta su quegli studi che mettono in luce la necessità di accordare il processo di elaborazione automatica del contenuto, nonché gli strumenti stessi di annotazione linguistica automatica che vi stanno alla base, alle specificità della lingua del diritto.

Portando l'attenzione su questo filone di ricerche, la Parte I si conclude passando in rassegna le attività di ricerca condotte nel campo dell'Intelligenza Artificiale e Diritto e finalizzate all'annotazione semantica di testi giuridici basata sull'annotazione linguistica automatica del testo[6]. È questo infatti il

[6]Vedi Paragrafo 2.3.2.2.

contesto in cui si colloca la metodologia di annotazione semantica di testi giuridici presentata nei Capitoli 7 e 6 di questo studio.

Parte II

L'obiettivo di questa seconda parte è quello di affrontare il primo ordine di questioni connesse con uno studio interdisciplinare e incrementale di testi giuridici. Si tratta delle "questioni di forma linguistica" per dirla con le parole di Garavelli (2001). Il modo in cui tali questioni sono state affrontate rappresenta uno degli elementi di originalità di questo lavoro.

Proponendosi di condurre uno studio dei rapporti tra lingua e diritto utilizzando strumenti di Trattamento Automatico del Linguaggio, si è ritenuto necessario prima di tutto verificare da un punto di vista sia quantitativo sia qualitativo l'affidabilità di tali strumenti nell'analisi di testi giuridici. Come messo infatti in luce nella Parte I, la questione è annoverata tra gli interessi di chi si basa su strumenti di annotazione linguistica automatica del testo come punto di partenza per analisi semantiche di diverso tipo. Inoltre, come discusso nel Paragrafo 3.1, determinare l'impatto che le caratteristiche linguistiche di un linguaggio specialistico hanno sui risultati dell'elaborazione linguistica automatica di testi di dominio è sin dagli anni '80 al centro degli studi della comunità di ricerca in linguistica computazionale.

Assumendo quest'ultimo tipo di studi come un punto di riferimento metodologico, nel Capitolo 3 il tema è affrontato in relazione all'uso di strumenti di Trattamento Automatico del Linguaggio basati su algoritmi di apprendimento automatico da dati testuali, strumenti che seguono cioè un "data–driven approach" per dirla con le parole di Nivre (2006). È questa infatti la tipologia di strumenti oggi più diffusa e che dimostra le migliori prestazioni di annotazione linguistica (come rilevato nelle più recenti campagne di valutazione di strumenti per l'analisi automatica del linguaggio natuale)[7].

Prendendo in esame un corpus di atti normativi italiani, linguisticamente annotati in maniera automatica fino al livello sintattico di analisi con la catena di strumenti di Trattamento Automatico del Linguaggio sviluppati dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa, e di tipo 'data–driven', l'attenzione è stata principalmente posta sull'impatto che la lingua del diritto ha sul grado di accuratezza dell'annotazione sintattica a dipendenze del corpus conside-

[7] Vedi Paragrafo 3.1.

rato. La scelta è legata alla centralità che tale livello di analisi ricopre per la successiva fase di annotazione semantica, costituendone l'imprescindibile punto di partenza.

La metodologia di indagine è basata sul confronto tra la qualità dei risultati dell'annotazione sintattica dei testi normativi e quella dell'annotazione di testi giornalistici, rappresentativi della lingua comune, presi come riferimento. Questo ha permesso *i)* di individuare aspetti specifici degli atti normativi che i criteri di annotazione linguistica automatica adottati per l'annotazione di testi giornalistici non coprono e di definire, di conseguenza, una serie di specializzazioni che riguardano più livelli di analisi (Paragrafo 3.3); *ii)* di quantificare l'accuratezza del livello di annotazione sintattica degli atti normativi, misurandola rispetto a diverse metriche di valutazione (Paragrafo 3.4); *iii)* di porre le necessarie premesse volte a definire una metodologia di adattamento di strumenti di Trattamento Automatico del Linguaggio all'annotazione sintattica di testi giuridici (Paragrafo 3.5).

Il secondo aspetto trattato nella seconda parte di questo studio riguarda le indicazioni, rilevanti ai fini di uno studio linguistico di testi giuridici, che si possono trarre dai risultati dell'annotazione linguistica automatica. L'obiettivo del Capitolo 4 è infatti quello di delineare le principali caratteristiche morfosintattiche, sintattiche e lessicali del corpus di atti normativi e amministrativi raccolto, sulla base della distribuzione di alcuni selezionati tratti linguistici rintracciati a partire dall'output dei vari livelli di annotazione automatica.

L'ottica comparativa è dunque l'elemento chiave delle analisi condotte, finalizzate, da un lato, a confrontare le caratteristiche linguistiche rintracciate nel composito corpus di testi giuridici raccolto con quelle rintracciate in due corpora di testi giornalistici, rappresentativi della lingua comune, presi come riferimento. In questo senso, l'obiettivo era quello di suggerire una possibile risposta 'operativa' alla dibattuta e aperta questione circa i non lievi problemi di delimitazione tra lingua del diritto e lingua comune. Dall'altro, essa è finalizzata a mettere a confronto tra loro le caratteristiche delle diverse tipologie di testi giuridici esaminati, affrontando la questione del carattere "multiforme e complesso" (Cortelazzo, 1997) della lingua del diritto.

La finalità era quella di dimostrare che *i)* i risultati ottenuti dall'annotazione linguistica automatica del testo, pur contenendo un margine di errore ulteriormente accentuato dalle specificità della lingua del diritto e dalle sue difficoltà di analisi, se appropriatamente esplorati sono un punto di partenza affidabile per definire una metodologia di monitoraggio linguistico articolato

su più livelli di analisi linguistica; *ii)* una tale strategia di monitoraggio consente di condurre indagini quantitative del profilo linguistico di testi giuridici, fornendo dimostrazioni empiriche di quanto fatto osservare negli studi linguistici tradizionalmente condotti con metodi manuali di indagine; *iii)* l'ottica comparativa adottata nel monitoraggio apre la strada a due scenari applicativi, pone cioè le basi per il futuro sviluppo di uno strumento a supporto delle attività di verifica della redazione 'chiara, semplice e comprensibile' di un atto normativo–amministrativo e di un indicatore del livello di leggibilità di testi giuridici basato sul monitoraggio linguistico.

Parte III

L'ultima parte di questo studio è interamente dedicata ad affrontare le questioni connesse con l'indagine del contenuto semantico–informativo di testi giuridici. In linea con l'approccio incrementale adottato, mentre sino a questo punto l'attenzione si è focalizzata sul livello di analisi relativo all'annotazione della struttura morfosintattica e sintattica dei testi giuridici, oggetto delle discussioni di questa terza parte è la metodologia di indagine adottata per accedere in modo incrementale all'informazione in essi implicitamente contenuta.

A questo scopo, nel Capitolo 5 sono prima di tutto passati in rassegna gli aspetti problematici da tenere in considerazione in un processo di accesso al contenuto di testi giuridici. Come chiarito nel Paragrafo 5.1, essi sono connessi con una ben nota peculiarità degli enunciati giuridici, caratterizzati dal "complesso intreccio di realtà giuridiche ed extragiuridiche"[8] che si riflette nel loro lessico, il loro riferirsi cioè nello stesso momento al mondo delle norme e a quello dei fatti regolati. Tema ampiamente discusso, ne sono qui considerate *i)* le conseguenze che ne derivano in un'ottica di rappresentazione formale del contenuto informativo in sistemi di organizzazione della conoscenza di dominio come le ontologie giuridiche[9]; nel caso in cui infatti le due componenti di realtà non siano tenute appropriatamente distinte, il rischio è quello di costruire ontologie affette da quella che Breuker e Hoekstra (2004) chiamano "epistemological promiscuity", riferendosi alla commistione di piani di organizzazione dell'informazione che caratterizza tali ontologie; e *ii)* le difficoltà con cui ci si scontra mettendo a punto approcci basati su di un'esplicita attenzione ai termini come principale via d'accesso al contenuto

[8] Belvedere (1994a, p. 23).
[9] Vedi Paragrafo 5.1.1.

testuale[10]. In questo caso, assumendo che i termini siano la prima istanza linguistica del contenuto, un tale intreccio di realtà, riflettendosi nella mescolanza in un testo giuridico di termini tecnico–giuridici (espressione della realtà giuridica) e di termini "fattuali" (espressione della realtà extragiuridica), mina la possibilità di individuare con certezza quale sia il lessico rilevante da cui partire per annotare il significato semantico–lessicale in esso contenuto e rappresentarlo in modo formale.

Tenendo in considerazione questi aspetti, nel Capitolo 5 sono inoltre descritti i passi necessari per accedere al contenuto dei testi giuridici e come essi si articolino in una successione incrementale. Sulla scia della riflessione di Buitelaar et al. (2005, pp. 3–12) riguardo al fatto che "terms are linguistic realizations of domain–specific concepts and are therefore central to further, more complex tasks", nel Paragrafo 5.2 è esposto il primo passo qui ritenuto fondamentale: l'identificazione e l'estrazione di terminologia dai testi. È dunque esposta la metodologia di estrazione automatica di terminologia da corpora testuali che sperimentata su un corpus di testi giuridici[11] si è rivelata affidabile per riuscire a individuare e fare distinzione tra termini tecnico–giuridici, termini fattuali e lessico comune.

Infine, nel Paragrafo 5.3 è presentato il secondo passaggio che consente di rendere espliciti aspetti sempre più complessi e avanzati del contenuto testuale. Esso è in linea, da un lato, con l'idea che in un processo completo di interpretazione di un testo giuridico sia necessario collocare il lessico ritenuto caratterizzante il testo "nel contesto degli enunciati", per dirla con le parole di Jori e Pintore (1995, p. 212); e che il significato di una parola, il 'concetto' di cui essa è espressione, sia unicamente definito sulla base dell'"insieme di regole che stabiliscono l'uso della parola"[12]. Dall'altro, esso abbraccia la visione di Charles Fillmore, per il quale "a language–internal semantic parsing of a sentence must be seen as merely a display of the lexical, grammatical and semantic material of the sentence"[13]. Essa, esprimendo i principi della 'Frame Semantics Theory', suggerisce come in una rappresentazione strutturata del significato lessicale sia necessario tenere in considerazione il contesto sintagmatico nel quale le parole occorrono, le proprietà semantico–combinatorie che permettono di renderne esplicito il significato.

[10] Vedi Paragrafo 5.1.2.
[11] Vedi Paragrafo 5.2.2.
[12] Bobbio (1976, p. 308).
[13] Fillmore (1985).

Sono questi i fondamenti teorici sui quali si basa la metodologia di annotazione semantica di testi giuridici descritta nel Capitolo 6 di questo studio e finalizzata a ricostruire il valore semantico dei rapporti sintagmatici tra i termini presenti in un testo. Essa prende le mosse dai principi organizzativi sottesi al progetto FrameNet basato sui presupposti teorici della 'Frame Semantics Theory' e il cui duplice intento è così descritto dai suoi creatori: "On this project our primary aim is to produce frame-semantic descriptions of lexical items [...] our concern with semantically tagged corpora is at both ends of our research"[14].

L'obiettivo è pertanto quello di mostrare come i principi di organizzazione e rappresentazione del significato lessicale di FrameNet lo rendano un modello particolarmente espressivo per l'annotazione semantica di testi giuridici, sia dal punto di vista strettamente linguistico sia dal punto di vista di strutturazione formalizzata del contenuto informativo. A questo scopo le potenzialità di FrameNet sono descritte mettendone a confronto i principi organizzativi con *i)* quelli di WordNet, il principale lessico computazionale oggi esistente basato su di un'organizzazione paradigmatica dello spazio semantico–lessicale delle parole, modello di riferimento dell'unico lessico computazionale giuridico oggi esistente: JurWordNet[15]; e con *ii)* i principi di organizzazione del significato sui quali si basano gli altri progetti esistenti, finalizzati a rendere esplicito il significato di una parola sulla base delle sue proprietà semantico–combinatorie[16].

Inoltre, il valore innovativo della scelta di prendere in considerazione FrameNet come modello di riferimento per l'annotazione semantica di testi giuridici è messo in luce passando in rassegna, nel Paragrafo 6.4, i vari usi e specializzazioni di dominio che dei modelli di rappresentazione del significato sono stati fatti sino ad oggi. Ne emerge che ben poca attenzione è stata dedicata al dominio giuridico. In particolare, come dichiarato da Rathert (2006), tranne rare eccezioni, "no one from *Frame Semantics* has ever looked for an application like forensic linguistics, and forensic linguistics have not yet started using the methods from computational linguistics". Per la lingua italiana l'unica eccezione è rappresentata appunto da JurWordNet, specializzazione del modello WordNet.

Nell'ultimo capitolo di questo studio (Capitolo 7), sono descritti gli adat-

[14]Lowe et al. (1997).
[15]Vedi Paragrafo 6.2.
[16]Vedi Paragrafo 6.3.

tamenti messi a punto per utilizzare FrameNet in una serie di esperimenti di annotazione semantica di un corpus di atti normativi statali, finalizzati a rendere esplicita ed organizzata l'informazione relativa a scenari deontici di 'obbligo', 'permesso' e 'divieto' in esso contenuti. Sebbene il tema della modalità deontica in enunciati giuridici sia stata a lungo studiata[17], in questo studio si intende restringere il campo all'analisi del modo in cui i frames presenti in FrameNet permettono di rappresentare in modo adeguato lo spazio semantico–lessicale di alcune delle principali unità lessicali, come ad esempio *obbligo*, *permettere*, *irrogare*, *divieto*, ecc..., presenti nel corpus in esame. A questo scopo è stata messa a punto una metodologia di annotazione che pur basandosi su quella proposta nel progetto FrameNet introduce alcune novità.

L'obiettivo è di dimostrare empiricamente le potenzialità di FrameNet, da un lato, come modello di rappresentazione del significato in grado di rendere esplicita la relazione tra semantica e realizzazione linguistica di elementi semanticamente rilevanti in un testo. Particolare attenzione è per questo dedicata a mettere in luce come alcune delle principali caratteristiche per lo più sintattiche individuate nel Capitolo 4, in fase di monitoraggio linguistico, influenzino il modo in cui il contenuto semantico è organizzato nel testo. In questo senso l'intento è quello di dimostrare come l'annotazione semantica si configuri come un ulteriore livello di annotazione testuale che si va ad aggiungere al precedente livello di annotazione sintattica a dipendenze. Dall'altro, l'obiettivo è quello di fornire una rappresentazione della conoscenza di dominio contenuta in un testo complementare a quella nota grazie a JurWordNet e alle ontologie giuridiche. Sono pertanto messi particolarmente in luce quegli aspetti innovativi che l'annotazione semantica basata sul modello FrameNet permette di rendere espliciti.

Gli obiettivi di ricerca

Gli intenti che hanno mosso l'intero lavoro di ricerca sono molteplici. Essi possono essere così riassunti nei seguenti punti:

- uno dei principali obiettivi è metodologico. Esso consiste nel dimostrare la novità e l'efficacia di un'analisi di testi giuridici basata sull'uso di strumenti di Trattamento Automatico del Linguaggio e articolata in più passi incrementali. Considerata la lunga tradizione degli studi sul

[17]Per una rassegna dei principali studi in materia e della loro centralità in uno studio linguistico di testi giuridici vedi Garavelli (2001, pp. 63–72).

rapporto tra lingua e diritto, nei capitoli che seguono particolare attenzione è dunque dedicata a mettere di volta in volta in luce l'apporto innovativo della particolare prospettiva linguistico–computazionale qui assunta;

- il secondo obiettivo è direttamente conseguente dall'analisi linguistica automatica presa come base dell'intero lavoro. Fondando le indagini sul profilo linguistico del corpus di testi giuridici raccolto sui risultati della loro annotazione linguistica automatica, l'obiettivo è quello di fornire una conferma quantitativa alle ricerche condotte in modo manuale dai linguisti, oltre a quello di mettere in luce nuovi aspetti che ricerche condotte con questo metodo di analisi non potevano portare alla luce;

- la terza finalità di questo lavoro di ricerca riguarda il nuovo modo di condurre un'analisi semantica di testi giuridici. In questo senso l'obiettivo consiste nel dimostrare empiricamente, grazie ad una serie di casi di studio, le potenzialità di una metodologia di analisi in grado di rendere espliciti quali siano all'interno di un corpus di testi giuridici *i)* i termini espressione dei principali concetti in esso presenti, *ii)* i tipi di relazioni semantico–contestuali tra termini, *iii)* il modo in cui è possibile rappresentare in modo formalmente strutturato il contenuto semantico–informativo veicolato da termini e relazioni;

- l'ultimo obiettivo riguarda le potenzialità applicative dei risultati di questo lavoro. Le future applicazioni spaziano tra diversi scenari che vanno *i)* dallo sviluppo di uno strumento a supporto dell'attività di 'drafting legislativo' che, rilevando in modo automatico le caratteristiche linguistiche del testo redatto, consenta al legislatore di verificare se e in che misura l'atto scritto soddisfi criteri di 'chiarezza, semplicità e comprensibilità' fino *ii)* alla definizione di metodi che, basati sull'uso di strumenti di Trattamento Automatico del Linguaggio specializzati per il trattamento della lingua del diritto, sono finalizzati ad estrarre e rendere espliciti quei nuclei di conoscenza implicitamente contenuti in corpora di testi giuridici che possono soddisfare i bisogni informativi dell'utente.

Parte I

Lingua e diritto: questioni dibattute e metodi di analisi

Capitolo 2

Il punto di vista di tre comunità di ricerca

È cosa nota che tradizionalmente comunità di ricerca diverse nei presupposti e negli intenti di ricerca hanno dimostrato un comune interesse per lo studio delle caratteristiche linguistiche dei testi giuridici. Linguisti, filosofi del diritto, giuristi e informatici interessati ad aspetti di formalizzazione del contenuto di documenti giuridici hanno annoverato tra i loro interessi di ricerca lo studio della lingua del diritto, mettendo in luce da più punti vista gli stretti rapporti tra aspetti di analisi linguistica del testo e aspetti di accesso, interpretazione e rappresentazione formale del suo contenuto informativo.

Più che nello studio di altri linguaggi espressione di domini di conoscenza specifici, come sottolinea Mortara Garavelli (2001, p. 4), "di fronte allo stesso oggetto di studio le pertinenze dei due campi, linguistico e giuridico, si intrecciano e si sovrappongono; e la reciprocità delle competenze si impone con una forza che non ha riscontri in quanto accade per gli altri linguaggi specialistici quando sono esaminati da chi li usa professionalmente e dai linguisti".

Il carattere interdisciplinare dell'oggetto di studio si riflette pertanto nella varietà di filoni di ricerca condotti da più comunità di ricerca, come è evidente consultando il ricco stato dell'arte redatto da Mortara Garavelli (2001), nella sezione del suo studio intitolata "Linguistica dei testi giuridici: tendenze attuali e prospettive di ricerca", che include:

- la semantica come "campo privilegiato"[1] dei cultori degli studi analitici

[1] Questa e le seguenti citazioni sono tratte da Garavelli (2001, pp. 34–54).

di filosofia del diritto;

- il lessico giuridico, ambito nel quale "è ovvio [...] prevalgono le competenze degli specialisti del diritto", anche se "è altrettanto scontato che queste debbano coesistere, e meglio se nello stesso studioso, con un ben fondato possesso di conoscenze storico–linguistiche, lessicologiche e lessicografiche";

- le ricerche linguistiche condotte dal punto di vista filologico e storico, che "hanno prodotto edizioni di testi antichi, indagini etimologiche e lessicologiche, raccolte bibliografiche, glossari";

- lo studio dell'italiano 'ufficiale' in stretta relazione con l'ottica della 'linguistica delle varietà', rispetto alla quale da un lato "i caratteri specifici dei testi giuridici vengono in primo piano" , dall'altro, in una prospettiva ribaltata, che vede "i testi giuridici al servizio della linguistica", i testi giuridici sono impiegati come fonti "per lumeggiare fenomeni linguistici di portata più ampia" di quella del dominio specialistico in questione;

- lo studio dei linguaggi settoriali e delle lingue speciali;

- studi di fonetica e sociolinguistica giudiziaria;

- lo studio della tecnica che ha per oggetto la redazione delle norme, cioè la 'legistica' o 'nomografia';

- la 'legimatica', disciplina nata dall'incontro fra legistica e informatica le cui attività di ricerca sono focalizzate "sul trattamento automatico dei testi giuridici e sull'applicazione dell'intelligenza artificiale al diritto";

- lo studio rivolto alla definizione di regole guida per la buona stesura di testi legislativi, "con proposte che vertono sull'organizzazione concettuale del testo, sulla semplificazione della sintassi e sul 'controllo' delle difficoltà lessicali".

La ricchezza del carattere multiforme di uno studio finalizzato a mettere in luce i rapporti tra lingua e diritto è anche al centro del recente studio di Visconti (2010, p. 8), che "tra i diversi, affascinanti, risvolti di tale questione, [...] ne privilegia uno: la riflessione sul rapporto tra il significato 'letterale' dei testi giuridici e la loro interpretazione".

Il lavoro qui presentato, differenziandosi dagli studi precedentemente condotti per la specifica prospettiva di osservazione assunta, prende tuttavia le mosse dagli aspetti di ricerca tenuti in considerazione fino a questo momento. In questo capitolo iniziale, in particolare, sono considerati e discussi quegli aspetti teorici e metodologici dei filoni di ricerca sopra menzionati che sono stati sino ad oggi maggiormente dibattuti. Essi costituiscono infatti il punto di partenza e l'orizzonte di riferimento di questo studio, rispetto ai quali i risultati delle analisi descritte nei capitoli successivi intendono portare il principale contributo innovativo.

Nei paragrafi che seguono sono pertanto messe al centro dell'attenzione le principali questioni dibattute e i metodi di analisi seguiti dalle tre maggiori comunità di ricerca interessate: quella dei linguisti (vedi Paragrafo 2.1), dei giuristi e filosofi del diritto (vedi Paragrafo 2.2) e quella di coloro che svolgono attività nell'ambito dei rapporti tra informatica e diritto, con particolare riguardo alle attività di più recente diffusione in materia di metodi e strumenti dell'intelligenza artificiale applicata al diritto (ambito noto come 'Artificial Intelligence and Law') e, in particolare, alle attività basate sull'utilizzo di metodi e tecniche di Trattamento Automatico del Linguaggio finalizzati all'annotazione semantica di testi giuridici (vedi Paragrafo 2.3).

2.1 Le attività di ricerca dei linguisti

"Della lingua giuridica, del suo lessico, dei suoi costrutti sintattici peculiari, della sua caratteristica testualità, delle condizioni della situazione comunicativa, sappiamo poco. Almeno dalla prospettiva dei linguisti. Perché, se sono numerosi gli studi sull'argomento da parte di giuristi, di filosofi del diritto, degli stessi operatori del diritto, gli scritti dei linguisti si contano (e non è un'iperbole) sulle dita di due mani". Così Michele Cortelazzo (1997) denunciava il fatto che lo studio del rapporto tra lingua e diritto è "un problema negletto agli studiosi di lingua". Secondo Cortelazzo è significativo il fatto che il nome che ricorre più frequente nelle sedi linguistiche è quello di Piero Fiorelli, accademico della Crusca, ma "prima di tutto un professore di materie giuridiche, non di linguistica o di storia della lingua".

Tuttavia, qualche anno più tardi, Francesco Sabatini (2003) riconosceva che, sebbene "non sono molti, a dir vero, i linguisti interessati agli usi giuridici del linguaggio", nonostante ciò, "negli ultimi tempi l'interesse è cresciuto e

si è articolato: all'attenzione riservata tradizionalmente agli aspetti lessicali e semantici si è aggiunta quella per gli aspetti sintattici e testuali".

Lontani dall'intenzione di proporre una nuova rassegna degli studi condotti dai linguisti in materia di 'lingua e diritto', ci si vuole qui concentrare piuttosto su due aspetti centrali per lo sviluppo delle successive fasi di ricerca, esposte nei capitoli che seguono. Tali aspetti riguardano:

- le riflessioni dei linguisti generate dalla difficoltà di definire il campo di indagine, di stabilire cioè cosa si debba intendere per 'lingua del diritto'. Secondo le riflessioni di Cortelazzo (1997), il carattere "multiforme e complesso" proprio del linguaggio giuridico è riconducibile da un lato alla varietà di tipologie di testi nei quali esso si instanzia, perché "più che in altri campi, sotto lingua giuridica intendiamo realtà spesso ben diverse: la lingua in cui si fissano i principi (testi legislativi), la lingua dell'attività giurisprudenziale (ordinanze, sentenze, ricorsi, memorie ecc.), la lingua usata dagli studiosi di diritto (nelle monografie, nei commenti a sentenze ecc.)"; dall'altro tale non unitarietà della lingua del diritto è dovuta agli stretti e biunivoci rapporti con la lingua comune e i linguaggi tecnico–specialistici, cioè alle sue "articolazioni orizzontali (per sottosettori del diritto)" e "verticali (con distinzioni fra espressioni puramente tecniche ed espressioni di uso comune)".

- le metodologie di indagine maggiormente seguite per descrivere le caratteristiche proprie della lingua del diritto.

La motivazione che ha portato a focalizzare l'attenzione su questi due aspetti di ricerca è duplice. In primo luogo, le discussioni intorno a questi due temi consentono di offrire una fondata base teorica delle metodologie di ricerca seguite in questo lavoro. In secondo luogo, tali temi sono centrali negli studi condotti sulla lingua del diritto, come dimostra il fatto che essi siano oggetto di dibattito condiviso da linguisti, giuristi e filosofi del diritto[2].

2.1.1 "La lingua del diritto ... dov'è?": aspetti teorici e metodologici

"Invidiabili cose sono i cataloghi di stelle o di piante o di composti chimici o di parti di macchine; per chi non soffra il pungolo dei problemi aperti, naturalmente. A quei livelli d'astrazione tutto si fa chiaro, e le lingue tecniche allo

[2]Vedi il Paragrafo 2.2.1.

stato puro si possono permettere la contemplazione d'una miriade di concetti e di termini collegati tra loro da perfette equivalenze, da corrispondenze biunivoche. La lingua del diritto non è fatta così".

Così Piero Fiorelli (1993) scrive nella Premessa all'"Indice della Lingua Legislativa Italiana (I.L.L.I.)", mettendo in guardia dalle difficoltà di redigere un catalogo esaustivo del lessico giuridico. Smentito dalla natura stessa della lingua giuridica sarebbe infatti "chi s'illudesse di mettere in fila tutti i termini giuridici d'un qualche testo, e quelli soli, separandoli dagli altri con un criterio rigido fissato una volta per tutte". Questo perché "gli usi lessicali del diritto non hanno limiti oggettivi assoluti: ne hanno, piuttosto, nel modo di considerare le tante facce della realtà e di comunicare questo modo"; perché, in ultima analisi, "il diritto ha bisogno d'indicare e di qualificare le cose più svariate e i loro più variati modi di essere".

È infatti questo compito del diritto di descrivere e regolare "le tante facce della realtà" la principale ragione dello stretto ma problematico rapporto tra lessico giuridico e lessico comune. Avverte Tullio De Mauro (1963, p. 426), affrontando la questione in una prospettiva diacronica: "un problema di rapporti tra usi linguistici correnti, non tecnici, e usi del linguaggio giuridico non si porrebbe nel caso di una codificazione che si preoccupasse in partenza di darsi una veste linguistica interamente formalizzata, ossia che si preoccupasse di muovere da una serie di esplicite definizioni relative ai termini adoperabili e alle loro regole d'uso. Una volta definiti formalmente, i termini di una codificazione del genere sarebbero sostanzialmente diversi dalle parole di uso comune [...] l'uso linguistico comune non avrebbe alcun peso nel determinare le scelte terminologiche e le regole d'uso, né vi sarebbe possibilità d'equivoco o di reciproca interferenza tra le formulazioni d'una codificazione siffatta e frasi del linguaggio comune eventualmente consonanti". Ma le cose sono andate diversamente. Secondo De Mauro (1963, p. 428)"i legislatori italiani hanno rinunziato ad una sistemazione rigida della terminologia ed hanno accettato, e consapevolmente, di operare nell'ambito dei valori lessicali risaputi".

Tale questione si inserisce nel ben noto dibattito circa il rapporto biunivoco tra linguaggi specialistici e linguaggio ordinario. È dunque riconducibile a quel fenomeno che Gian Luigi Beccaria (1973) definisce "escursione terminologica", indicandolo come caratteristica peculiare di ogni linguaggio specialistico. L'allusione è al fatto che "tra vocabolario comune e vocabolario tecnico si ergono sempre più esili barriere", conseguenza della "crescente forza espansiva" e del "prestigio reale nell'uso parlato e scritto di cui sono

dotati i linguaggi settoriali".

Ma fino a che punto la lingua del diritto è una lingua con un "vocabolario tecnico"? È interessante confrontare le risposte che provengono da rappresentanti della comunità dei linguisti, dei giuristi e da parte di chi, pur da linguista, si è dedicato allo studio della lingua del diritto in costante dialogo con i giuristi e i filosofi del diritto. In qualità di linguista ne scrive Bice Mortara Garavelli (2001): quando si parla di lingua del diritto si ha a che fare con una varietà di lingua che differisce "dalla matrice comune per l'impiego di tecnicismi lessicali e per una formalità di registri che è altra cosa dalla formalizzazione delle lingue speciali scientifiche". Ne scrive il giurista Sabino Cassese (1992): "Se è certo che vi sia una linguistica giuridica, è dubbio che vi sia un linguaggio giuridico, separato da quello comune alla stregua dei linguaggi formali e simbolici delle *hard sciences*. In realtà, il linguaggio giuridico è un sottinsieme, distinto ma non separato dal linguaggio generale o comune, con alcuni termini propri (che sono pochi, ed hanno – per lo più – una doppia appartenenza, al linguaggio giuridico e a quello comune [...]) e senza una propria sintassi, anche se, proprio per essere distinto, è percorso da tensioni che lo differenziano dagli usi informali e quotidiani di una lingua". Ne scrive Piero Fiorelli (2008)[3]: "Il fatto è che il diritto è qualcosa di straordinariamente esteso, da non paragonare ai settori dell'operare o del sapere che sono oggetto della maggior parte delle lingue tecniche. L'esperienza umana tutta intera rientra nel diritto, così come per un altro verso rientra nel linguaggio. [...] Così, alla lingua del diritto la qualifica di lingua tecnica sta un po' stretta".

Dalle loro risposte se ne deduce un consenso unanime nel riconoscere come il rapporto tra lingua del diritto e lingua comune non sia il rapporto che occorre tra un linguaggio tecnico–specialistico e il linguaggio ordinario, perché la lingua del diritto non è lingua tecnica, separata dal linguaggio comune.

Si interroga allora Fiorelli (2008): "La lingua del diritto [...] dov'è? [...] in una classificazione delle lingue tecniche che miri a essere rigorosa ed esaustiva, viene il dubbio se le si debba riconoscere un posto a sé, o più posti in rapporto ai suoi diversi livelli (dal legislativo al giudiziario, dal diplomatico al commerciale, dal notarile all'amministrativo ...), o ancora un posto diverso da tutti gli altri".

[3]Le citazioni da Fiorelli (2008) sono tratte dalla sezione intitolata "Qualche dubbio sulla lingua del diritto".

Da un punto di vista lessicale, la risposta è la seguente: la lingua del diritto è pienamente identificabile *i)* nelle 'ridefinizioni', risultato di ciò che Mortara Garavelli (2001, p. 11) definisce "l'atto (o, se si preferisce, il gioco) linguistico del 'ridefinire"', "il riuso specialistico di termini del linguaggio ordinario", e *ii)* nei tecnicismi, ciò che Fiorelli (2008) chiama "un'insalata, una mescolanza di tecnicismi".

Tuttavia, le ricerche circa la definizione dello statuto della lingua del diritto non si esauriscono nel dibattito sul suo rapporto con la lingua comune. È anche in questione il rapporto con i linguaggi tecnico–specialistici oggetto del discorso giuridico. Fanno parte integrante della lingua del diritto anche le "articolazioni orizzontali (per sottosettori del diritto)" messe in luce da Cortelazzo (1997). Come sottolinea Mortara Garavelli (2001, p. 24), "ciò che si intende per linguaggio giuridico (o legale) non è dedicato esclusivamente all'esame di questioni di diritto; anzi, una parte cospicua dei testi giuridici che hanno effetti applicativi è occupata dall'esame di eventi e situazioni concrete senza il cui accertamento non è possibile individuare le norme applicabili. Il giudice che discute sulla validità delle prove offertegli, il notaio che descrive l'immobile compravenduto, il presidente della regione che stipula un contratto di appalto, i privati che redigono un contratto parlano o scrivono per esporre situazioni di fatto e dati concreti senza i quali gli effetti giuridici sarebbe inesistenti o viziati".

Anche definire cosa si debba intendere per morfosintassi e sintassi della lingua del diritto è un compito che presenta alcune difficoltà. È interessante qui far osservare come sia proprio rispetto a questo livello di analisi linguistica che sono state (sino ad oggi) applicate nuove metodologie di indagine.

Tra tutte, la prospettiva metodologica assunta da Giovanni Rovere (2005) è la più stimolante per le ricerche condotte in questo studio. Assumendo un'ottica comparativa, Rovere parte dall'assunto teorico, esposto nei "manuali di linguistica [...] che la morfosintassi delle lingue speciali non si caratterizza per la presenza di tratti esclusivi, ma per la selezione, significativa da un punto di vista quantitativo, di alcune fra le opzioni offerte dal sistema della lingua comune".

Nella sezione di "Premesse metodologiche" del suo studio egli, mettendo in evidenza le potenzialità di ricerche linguistiche condotte a partire da collezioni documentali, sottolinea che "la linguistica dei *corpora* svolge [...] una funzione di documentazione, a cui può aggiungersi una ponderazione dei singoli dati, tesa a evidenziare ciò che, dal punto di vista della frequenza, è centrale e ciò che è invece marginale". Tale approccio è basato su due

presupposti metodologici centrali nello studio qui condotto.

Da una parte, vengono messi in evidenza i vantaggi di un'analisi finalizzata a documentare gli usi linguistici concreti sulla base di evidenze testuali. È la direttiva di ricerca che auspica Mortara Garavelli (2001). Tra i possibili sviluppi degli studi sulla lingua del diritto, l'autrice immagina che il "campionario di esempi" da lei raccolto possa "servire come indicazione di spunti per lavori che daranno le necessarie conferme quantitative, se condotti sistematicamente con i metodi della 'linguistica dei *corpora*"'. Che i suoi suggerimenti metodologici siano stati raccolti, lo dimostra non solo lo studio di Rovere, ma anche quello di Patrizia Bellucci (2005), dove l'autrice conduce un'analisi delle caratteristiche della lingua giudiziaria a partire dallo studio delle varie tipologie di testi/discorsi (intercettazioni, interrogatori, dibattimenti) prodotti durante il processo.

Dall'altra, l'allusione di Rovere alla "ponderazione dei singoli dati" fa riferimento ad una metodologia di analisi comparativa, considerata particolarmente produttiva nelle ricerche basate su corpora finalizzate a mettere in luce le specificità di varietà linguistiche. L'assunto di base consiste nel rintracciare le caratteristiche proprie di una varietà a partire dalle dimensioni di variazione più significative osservate nel confronto con il codice linguistico di riferimento.

La soluzione operativa proposta da Rovere all'aperta questione circa le difficoltà di definire i confini morfosintattici e sintattici della lingua del diritto consiste infatti nell'individuarne le peculiarità analizzando in modo comparativo alcuni tratti linguistici presenti in diverse tipologie di testi giuridici e in testi giornalistici, assunti come rappresentativi della lingua comune.

Inoltre, "la necessità di sviluppare procedure atte a identificare il valore tecnico di elementi che nei manuali e nei dizionari specializzati non sono sottoposti a definizione esplicita", riconosciuta da Rovere, è un'ulteriore questione metodologicamente rilevante. Il riferimento è al fatto che le variazioni nella distribuzione di caratteristiche morfosintattiche e sintattiche rintracciate nei corpora rappresentativi della lingua del diritto e quella ordinaria sono indicative e rilevanti per uno studio semantico di testi giuridici[4].

[4]Punto di osservazione privilegiato delle analisi di Rovere (2005) è lo studio del verbo, condotto mettendo a confronto il quadro valenziale dei verbi che occorrono in testi giuridici e in testi giornalistici. Nello studio da lui condotto, l'autore trova conferma del fatto che, mentre "nelle lingue speciali la configurazione tecnica degli argomenti permette in genere distinzioni alquanto nette tra i significati tecnici del verbo", al contrario nella lingua comune "le varianti contestuali comportano spesso soluzioni sfumate".

2.2 Le attività di ricerca dei giuristi e dei filosofi del diritto

"La necessità di porsi questioni linguistiche in stretta connessione con questioni giuridiche ha impegnato da tempo i giuristi, sui versanti teorico e applicativo: e sul primo versante principalmente, anche se non esclusivamente, i cultori degli studi analitici di filosofia del diritto. Campo privilegiato, la semantica, per ragioni evidenti". Così Bice Mortara Garavelli (2001, p. 34) inizia la sezione intitolata "Linguistica dei testi giuridici: tendenze attuali e prospettive di ricerca", suggerendo come lo studio della lingua del diritto sia questione che trascende i confini prettamente linguistici.

Anche se non in modo esclusivo[5], l'attenzione per lo studio della lingua del diritto è principalmente riconducibile agli interessi di giuristi e filosofi facenti capo alla scuola analitica italiana di filosofia del diritto. Tra i suoi fondatori nel secondo dopoguerra insieme a Norberto Bobbio, Uberto Scarpelli (1969) ricorda infatti che i filosofi del diritto di indirizzo analitico "hanno posto al centro della loro attenzione il linguaggio giuridico, soprattutto sotto profili semantici, di qui sono risaliti a temi linguistici ed in specie semantici di ordine generale".

Non è negli intenti di questo studio quello di fornire una rassegna dei contributi al dibattito analitico in ambito giuridico in Italia[6]. Ci si vuole qui piuttosto limitare a ricordare come nelle intenzioni dei promotori dell'approccio analitico italiano l'attenzione alla lingua del diritto, in quanto carattere intrinsecamente costitutivo l'oggetto d'indagine filosofica, serva a proporre una soluzione e a superare alcuni problemi teorici interni alla disciplina.

Tale approccio costituisce infatti una svolta rispetto alla prospettiva d'indagine di tutti quei giuristi che tendono a tralasciare il fondamento linguistico–semantico delle loro teorie. Secondo l'analisi di Scarpelli (1969), essi infatti "affrontano in genere le questioni semantiche, che si aprono nel loro lavoro, in prospettive piuttosto ristrette, senza allargare l'orizzonte oltre l'universo della cultura giuridica". Ma soprattutto, "i giurisisti [...] considerano ed usano il linguaggio come uno strumento semplice ed onesto, intorno a cui

[5]Vedi per un approfondimento di come l'interesse filosofico alla lingua del diritto non sia una prerogativa analitica la raccolta curata da Scarpelli e Di Lucia (1994).

[6]Per una rassegna dei maggiori temi di dibattito vedi Scarpelli (1976) e Scarpelli e Di Lucia (1994).

non c'è troppo da discutere, perché quanto alle sue finalità ed al suo impiego non possono nascere gravi dubbi".

Al contrario, l'approccio analitico è caratterizzato dal monito all'attenzione linguistica che un giurista deve dedicare durante le proprie attività, attività che, secondo Scarpelli (1969), "riguardano il linguaggio ed hanno come strumento il linguaggio", perché "se c'è un'attività che richieda una consapevolezza linguistica, questa è l'attività dei giuristi".

È questa la prospettiva di indagine che risulta essere particolarmente significativa ai fini di questo lavoro. In quest'ottica, infatti, partendo dal presupposto che il diritto è composto principalmente di segni linguistici, il giurista svolgendo la sua tradizionale attività, quella cioè di interpretazione della legge, non compie altro che un compito di semiotica linguistica. Si interroga infatti retoricamente Norberto Bobbio (1976, p. 306): "Che altro è [...] l'interpretazione della legge se non l'analisi del linguaggio del legislatore, cioè di quel linguaggio in cui vengono espresse le regole giuridiche?"[7].

In particolare, come ricordano Jori e Pintore (1995, p. 205), la concezione che l'interpretazione debba essere "intesa come l'identificazione delle norme giuridiche, cioè del significato degli enunciati normativi giuridici" si basa su due presupposti imprescindibili: "ovviamente presuppone che il diritto sia composto di norme intese come significati e che questi significati vadano ricavati da specifici e individuabili enunciati".

L'indicazione di come poi collegare il significato con gli enunciati normativi viene, secondo gli autori, proprio dall'articolo 12 "Interpretazione della legge" del Codice Civile[8], dove si prescrive di applicare (interpretando) la

[7]Come spiega Bobbio (1976, pp. 311-313), l'attacco polemico è contro la concezione sino ad allora invalsa di interpretazione giuridica, comunemente intesa come "un procedimento intellettivo che [...] permette di guardare al di là delle proposizioni, che [...] permette di aprire, per così dire, una finestra attraverso le proposizioni per vedere che cosa c'è dietro, [...] di saltare al di là del linguaggio". Ma seguendo questa concezione "cadrebbe la riduzione della giurisprudenza ad analisi del linguaggio". Andare infatti alla ricerca di "ciò che sta al di là delle proposizioni normative" significa ricercare "qualcosa che non sia del tutto riconducibile alle proposizioni stesse", un 'qualcosa' che viene comunemente chiamato "spirito, volontà, pensiero, intenzione del legislatore". L'interpretazione di questo 'qualcosa', l'interpretazione cioè della *mens legis*, pur rappresentando qualcosa di diverso dall'interpretazione della lettera, l'interpretazione cioè dei *verba*, deve essere condotta con gli stessi mezzi: "Per interpretazione dell'intenzione, insomma, si deve intendere l'uso di tutti quei mezzi che sono atti a stabilire il significato di una parola o di un gruppo di parole usate: ma tutti questi mezzi, si ricordi, sono linguistici".

[8]L'articolo recita così: "Nell'applicare la legge non si può ad essa attribuire altro senso che quello fatto palese dal significato proprio delle parole secondo la connessione di esse, e

legge guidati dalla "connessione" delle parole in essa contenute.

Ciò suggerirebbe, secondo Jori e Pintore (1995, p. 209), che "i problemi di significato degli enunciati giuridici possono essere affrontati solo risolvendone i problemi sintattici". In questo modo verrebbe anzi espressamente riconosciuta una priorità alla risoluzione dei problemi sintattici su quelli semantici ai fini dell'interpretazione della legge. Chiosano gli autori: "possiamo dire che i problemi sintattici vanno risolti 'prima' di quelli lessicali e semantici in senso stretto". L'attenzione si avvicina a quella del linguista che svolge la propria analisi linguistica di un testo articolandola nei diversi livelli di analisi: "Tra i primi problemi spiccano, nella interpretazione giuridica, le difficoltà di accertamento della struttura sintattica degli enunciati, per esempio di comprensione del significato della punteggiatura o del particolare ordine delle parole in enunciati spesso assai complessi e spesso poco curati da questo punto di vista 'formale' (che in questo caso vuol dire appunto sintattico–grammaticale)". Vengono poi i problemi semantici, problemi cioè che riguardano soprattutto questioni di "accertamento e attribuzione di significato [...] ai singoli termini del discorso giuridico".

In altre parole, Jori sottolinea il fatto che "non può essere accolta la concezione [...] che l'interpretazione del diritto consista semplicemente nel sommare le interpretazioni delle singole parole dei discorsi giuridici"[9]. Al contrario, l'attività interpretativa deve articolarsi nei diversi livelli di analisi linguistica, sintassi, semantica e pragmatica. Essa cioè "richiede contemporaneamente: una attenta considerazione della struttura sintattica e grammaticale; una comprensione del suo lessico; una collocazione di questo nel contesto degli enunciati e delle unità di discorso maggiori, dai commi e articoli di legge o singole sentenze, all'intera regolamentazione della disciplina e alla linea giurisprudenziale rilevante; una serie di complesse e delicate considerazioni pragmatiche, cioè riguardanti i possibili effetti generali delle varie possibilità interpretative, la valutazione delle situazioni in cui queste possibilità si possono collocare, in quanto rilevanti alla determinazione stessa del significato"[10].

---

dalla intenzione del legislatore". Non è qui intenzione riproporre il dibattito nato intorno alle diverse letture di questo articolo del Codice Civile. Se n'è fatto riferimento in questo contesto riportando la proposta di lettura suggerita da Jori e Pintore (1995). Per una discussione dei diversi punti di vista e delle questioni interpretative coinvolte si rimanda a quanto esposto da Belvedere Belvedere (2000).

[9] Jori e Pintore (1995, p. 212).

[10] Jori e Pintore (1995, pp. 212-213).

2.2.1 Questioni di lessico

Ricorda Giovanni Tarello (1976, p. 377): "l'attenzione che l'approccio analitico pone sul linguaggio si rivolge anche e particolarmente al lessico". È infatti dal lessico che, secondo Bobbio (1976, p. 308), inizia l'attività di analisi linguistica del giurista analista impegnato a determinare il "significato delle parole che entrano a far parte della proposizione normativa o del gruppo delle proposizioni normative che formano oggetto della sua ricerca". Tale attività consiste principalmente in un costante processo di definizione, dal momento che, come ricorda Scarpelli (1976b), nel discorso giuridico "il significato di una parola non è qualcosa che sia intrinsecamente e definitivamente legato ad essa, ma dipende soltanto dalle regole che per l'uso di quella parola valgono in quel determinato sistema di linguaggio".

È qui interesse soffermarsi proprio su questo monito analitico all'**uso**, che ha guidato gli studi indirizzati alla ricerca di quale sia lo statuto del lessico giuridico soprattutto nei suoi rapporti con l'italiano comune. Essi affrontano infatti la questione in modo diverso da quanto fatto dai linguisti[11].

In primo luogo, assumendo una prospettiva diacronica, la questione non riguarda tanto, come ritiene De Mauro (1963), un problema di "scelta" ad opera del legislatore che ha "rinunziato ad una sistemazione rigida della terminologia". La difficoltà di definire i confini tra il lessico della lingua del diritto e quello della lingua comune riguarda piuttosto la natura specifica della lingua del diritto, "frutto di una secolare opera di ricostruzioni parziali all'interno dei linguaggi naturali, ricostruzioni parziali incidenti principalmente sulla dimensione semantica dei linguaggi stessi", come ricorda Scarpelli (1969). Nata dunque nell'alveo del linguaggio comune, la lingua del diritto diventa "un linguaggio tecnico, nel senso soprattutto, di un vocabolario tecnico introdotto nella struttura di un linguaggio naturale".

È dunque sulla struttura del linguaggio naturale che si innesta l'attività di 'costruzione giuridica', intesa come attività di costruzione di concetti a partire dall'uso delle parole nella legge. Pertanto, "l'insieme delle regole che stabiliscono l'uso di una parola costituisce il concetto corrispondente a quella parola. Il concetto di proprietà, di mandato, di mutuo e simili, è dato dall'insieme delle regole che stabiliscono l'uso della parola mandato, proprietà, mutuo e simili" (Bobbio, 1976, p. 308).

In quest'ottica, è allora l'uso a determinare i modi e le forme del rapporto tra lingua del diritto e lingua comune, nel senso che, come afferma Andrea

[11] Vedi Paragrafo 2.1.1.

Belvedere (1994b, p. 405), "le differenze tra linguaggio ordinario e giuridico non riguardano i termini, ma il loro uso, nel senso che uno stesso termine potrà ricorrere negli enunciati di entrambi i linguaggi, ma con significati (più o meno) differenti". Ciò determina di conseguenza l'interesse degli studi di semantica giuridica per le **definizioni** dei concetti giuridici attraverso un'attenta analisi delle loro concrete regole d'uso[12].

Un tale approccio allo studio del lessico giuridico rende possibile una classificazione delle tipologie di termini che occorrono nei testi giuridici. È quanto è stato fatto da Scarpelli (1976b) e successivamente da Belvedere (1994a e 1994b). Ai fini dello studio qui condotto, la classificazione proposta da Belvedere (1994a) è la più significativa.

Egli parte dal presupposto che "nel lessico giuridico [...] si riflette un complesso intreccio di realtà giuridiche ed extragiuridiche"[13]. Le prime fanno riferimento al 'mondo delle norme', le seconde al 'mondo dei fatti'. Come si può vedere nella Tabella 2.1, dove è riassunta schematicamente la classificazione da lui proposta, questa distinzione gli permette di individuare, sulla base del diverso tipo di realtà di riferimento, quali siano le diverse tipologie di termini che fanno parte del lessico giuridico.

Una tale prospettiva è qui considerata di grande interesse e costituisce infatti l'orizzonte teorico dell'intero approccio all'accesso al contenuto informativo di testi giuridici messo a punto in questo lavoro. La classificazione di Belvedere rappresenta, in particolare, il punto di partenza *i)* della metodologia di estrazione automatica di terminologica da testi descritta nel Paragrafo 5.2.1, finalizzata a distinguere in modo automatico le diverse tipologie di termini presenti in corpora di testi giuridici e *ii)* dell'approccio all'annotazione semantica descritto nel Capitolo 7, finalizzata a rendere esplicito il modo in cui le due principali componenti semantiche presenti in un periodo giuridico interagiscono tra di loro.

Inoltre, come per i linguisti anche per i filosofi del diritto le attività di studio del lessico giuridico non si esauriscono nell'analisi dei rapporti tra la lingua del diritto e la lingua comune. Ciò è riconducibile al fatto che, come ricorda il giurista Giuseppe Zaccaria (2003), il discorso giuridico è composto non solo da due, ma da "tre polarità linguistiche: la lingua comune, utilizzata dalla totalità dei parlanti, la lingua speciale del diritto e altre lingue speciali,

[12]Per un approfondimento del tema delle 'definizioni' si rimanda ad alcuni dei contributi più significativi quali Belvedere (1994a,b, 1998), Jori (1994) e Scarpelli (1976b, 1959).

[13]Belvedere (1994a, p. 23).

Tipologia di termini	Esempi
Tipo di realtà di riferimento: **realtà giuridica**	
"termini indicanti norme giuridiche"	*legge, disposizione, norma, ordinamento giuridico*
"termini attraverso i quali si esprime sul piano linguistico la funzione prescrittiva delle norme, che qualificano giuridicamente comportamenti"	*potere, dovere, obbligo, diritto, facoltà, divieto, vietato, obbligatorio, lecito, permesso*
Tipo di realtà di riferimento: **realtà extragiuridica**	
termini "fattuali" che "nel loro significato non contengono riferimenti a norme giuridiche"	*fumo, muro, concime, siepe*
termini "normativi" il cui "(corretto) uso richiede un previo accertamento del rapporto esistente tra una realtà di fatto ed una o più norme giuridiche [...] in cui vengono esplicitate le condizioni d'uso del termine"	*figlio legittimo, contratto, assenza*

Tabella 2.1: Tipologie di termini del lessico giuridico proposte da Belvedere (1994a).

dipendenti da settori di conoscenze o sfere di attività specialistici". In una prospettiva di semiotica giuridica, ciò comporta che, oltre ai problemi connessi con le difficoltà di stabilire i modi e le forme dello stretto rapporto tra linguaggio giuridico e linguaggio ordinario, "un secondo ordine di problemi deriva dai rapporti tra il diritto e i linguaggi tecnici e specialistici". Anche in questo caso, tale stato di cose è determinato dalla natura stessa del diritto. Ricorda in proposito Scarpelli (1959): "Quando, nel caso del diritto, si tratta di dar norma alla vita comune e ad attività specialistiche di ogni genere in mille diversi aspetti, è necessario disporre della ricchezza del linguaggio comune e dei vari linguaggi specialistici: il linguaggio tecnico della disciplina normativa può integrare quei linguaggi, costituirà la struttura intorno alla quale se ne organizzerà l'impiego, ma di quei linguaggi non si può fare a meno".

In particolare, i problemi connessi con l'interpretazione di quei termini che appartengono a linguaggio tecnico–specialistici sono dovuti a questioni di definizione. Quando il diritto assume "un termine in un suo significato tecnico, quello che esso possiede in una disciplina scientifica non giuridica (medicina, psichiatria, scienza alimentare, eccetera)" l'impegno definitorio deve essere maggiore, avvertono Jori e Pintore (1995, pp. 210–211). Nel caso di termini assunti da un linguaggio specialistico, infatti, precisano, "il

significato sarà di solito assai più precisamente delimitato e il diritto potrà fino a un certo punto riuscire a liberarsi del riferimento al significato comune e ai presupposti in esso incorporati". Ciò comporta di fatto un problema di attribuzione del compito di interpretazione giuridica, portando a "chiederci quanto l'interpretazione delle norme giuridiche in cui l'enunciato è usato debba essere compiuta da un esperto della disciplina e quanto dall'esperto di diritto".

2.3 Le attività di ricerca in Informatica e Diritto

L'analisi delle ricerche condotte nell'ambito degli studi informatico–giuridici è finalizzata a mettere in luce le sfide tutt'ora aperte connesse con la necessità di sviluppare metodi e sistemi di gestione automatica di testi giuridici sulla base del loro contenuto informativo, a partire dal trattamento automatico della lingua in cui i testi sono scritti. Le discussioni condotte in questo paragrafo sono pertanto finalizzate a delineare alcuni dei posssibili scenari applicativi aperti da questo studio.

Data la vastità del tema, è importante chiarire che i due principali filoni di ricerca in informatica giuridica sui quali ci si intende soffermare riguardano:

- le attività di ricerca legate alla 'legimatica', le applicazioni cioè dell'informatica a supporto della 'legistica' come ausilio sia durante la fase ex–ante di corretta redazione di un testo legislativo sia durante la fase ex–post di controllo e gestione del testo redatto;

- le attività condotte in materia di intelligenza artificiale e diritto ('Artificial Intelligence and Law') e finalizzate a mettere a punto metodi di rappresentazione formale della conoscenza di dominio contenuta in varie tipologie di corpora di testi giuridici.

L'interesse è in particolare circoscritto a quelle attività che in questi due ambiti hanno messo a punto metodologie finalizzate allo sviluppo di sistemi legimatici e sistemi in grado di svolgere compiti di gestione della conoscenza giuridica utilizzando strumenti di Trattamento Automatico del Linguaggio.

Il paragrafo è dunque così organizzato: nel Paragrafo 2.3.1 sono delineate le principali attività di ricerca nell'ambito della legimatica che hanno rivolto

particolare attenzione alle difficoltà e alle sfide che le caratteristiche linguistiche dei testi giuridici pongono nello sviluppo di applicazioni informatico–giuridiche. In particolare, nel Paragrafo 2.3.1.1 sono passate in rassegna le attività nate dal dibattito sulla legimatica e finalizzate al trattamento automatico della documentazione giuridica realizzato con metodi e tecniche di Trattamento Automatico del Linguaggio. Tenuto conto del fatto che, come afferma Mercatali (2004), "si può dire che la culla della legimatica sia stata l'Italia", sono passate in rassegna attività di ricerca unicamente italiane.

Al contrario le attività di ricerca condotte nell'ambito degli studi in intelligenza artificiale e diritto ('Artificial Intelligence and Law', da ora in poi AI&Law) hanno storicamente radici nord–americane. Pertanto, la rassegna presentata nel Paragrafo 2.3.2 mira a riportare il dibattito a livello internazionale. Anche in questo caso l'attenzione è posta sulle attività che in quest'ambito fanno affidamento su strumenti di Trattamento Automatico del Linguaggio per sviluppare metodi di gestione della conoscenza contenuta in collezioni di testi giuridici.

2.3.1 "Legimatica: informatica per legiferare"

Disciplina nata all'inizio degli anni '90 dall'incontro tra 'legistica' e informatica, la 'legimatica' si caratterizza per "approccio interdisciplinare complesso", che mira a "applicare l'informatica alle tecniche legislative, intese soprattutto come valutazione *ex ante* ed *ex post* dell'efficacia di una legge" e che "sperimenta l'intersezione di molteplici discipline informatiche come supporti a più tecniche legislative. Si colloca quindi ad un crocevia tra teoria generale del diritto, informatica tradizionale, intelligenza artificiale, linguistica e scienza cognitiva"[14].

In Italia le ricerche nell'ambito della legimatica sono infatti realizzate da ricercatori che afferiscono a due centri di ricerca complementari: l'IDG-

[14]Così Taddei Elmi (1995). Nella definizione data da Mercatali (1995), a cui si fa tradizionalmente riferimento, la legimatica viene infatti così descritta: "La legimatica si occupa della modellizzazione del ragionamento e delle procedure relative alla produzione legislativa, quindi della redazione dei testi legislativi (studio ora prevalente), dell'attività politico–decisionale, dell'analisi di fattibilità, della verifica d'efficacia e così via. Si rifà alla teoria normativa del diritto, utilizza metodologie logiche, linguistiche e pragmatiche (in particolare le tecniche legislative) per l'analisi dei testi normativi. Ha per scopo l'informatizzazione del processo di produzione normativa. Si propone di offrire conoscenze e strumenti informatici alle assemblee legislative e più in generale a tutti i produttori di norme".

ITTIG, Istituto per la documentazione giuridica del CNR di Firenze (dal 2002 Istituto di Teoria e Tecniche dell'Informazione Giuridica)[15] e il CIRSFID (Centro Interdipartimentale di Ricerca in Storia del Diritto, Filosofia e Sociologia del Diritto e Informatica Giuridica dell'Università degli Studi di Bologna)[16].

Agli scopi dello studio qui condotto, è importante soprattutto chiarire che l'attenzione con cui i primi sviluppatori di strumenti legimatici guardavano al linguaggio del testo normativo era per lo più mossa da un interesse di controllo ex–ante del testo redatto.

La situazione è chiaramente delineata da Cassese (1992), il quale denuncia il fatto che, sebbene la legistica abbia un carattere preminentemente linguistico, avvalendosi "dei principi di linguistica giuridica e di semiotica giuridica", tuttavia essa è stata storicamente considerata sia da giuristi sia da linguisti una mera tecnica. Da un lato, infatti, secondo Cassese, i giuristi non si danno cura della fase di scrittura delle norme, concentrandosi piuttosto sull'interpretazione; dall'altro, la linguistica giuridica, riducendo "la scienza della legislazione a tecnica di redazione delle norme", finisce per fornire unicamente indicazioni prescrittive, dettando per lo più "requisiti, come quelli di unità (presenza di una idea fondamentale), completezza (pieno svolgimento del tema di fondo), coerenza (non contraddittorietà dei concetti esposti), coesione (buon collegamento delle parole che costituiscono il testo)".

In Italia, le prime attività legimatiche avviate agli inizi degli anni '90 si sviluppano infatti non a caso negli anni in cui inizia a nascere la consapevolezza per una maggiore attenzione alla redazione di testi giuridici (e amministrativi) scritti in una lingua chiara e comprensibile. È quanto sottolinea De Mauro, ricordando come in Italia, sulla scia di quanto stava avvenendo sul piano internazionale, "lo sviluppo del dibattito sulla tecnica legislativa ha messo in luce come anche il legislatore, nei momenti in cui crea la norma, debba tener conto del modo in cui viene espressa e ricevuta"[17]. È questo il contesto in cui sono infatti avviate in questi anni le prime iniziative istituzionali finalizzate alla stesura di manuali e codici di regole e suggerimenti per la redazione di atti normativi e amministrativi scritti in un linguaggio chiaro, semplice e comprensibile [18].

[15] http://www.ittig.cnr.it/

[16] http://www.cirsfid.unibo.it/CIRSFID/default.htm

[17] La citazione è tratta dal contributo di De Mauro a (Zuanelli, 1990, p. 219).

[18] Vedi Piemontese (1999, p. 270–271) per una rassegna delle tappe più significative fino alla fine degli anni '90 "segnate dall'apparato statale nella direzione della dichiarazione

Nell'ambito di una tale attenzione alla qualità del linguaggio da usarsi in fase di redazione, le prime attività legimatiche sono dunque finalizzate a mettere a punto metodi e tecniche informatici in grado di offrire "una serie di strumenti che vanno dai semplici editori di testi con correttori ortografici, ai controlli di leggibilità [...] e alle tecniche di disambiguazione appoggiate su approcci di intelligenza artificiale" (Taddei Elmi, 1995, p. 271)[19]. È in quest'ottica che si delineano fruttuosi rapporti di scambio reciproco tra legistica e informatica, in base ai quali "l'informatica mette a disposizione della legistica nuovi algoritmi, nuovi strumenti" e la legistica scopre "l'informatica come strumento idoneo a gestire i modelli che essa ha prodotto" (Mercatali, 1995, p. 39).

È allora che iniziano a diffondersi in Italia le primissime attività volte all'applicazione di metodi quantitativi per la misurazione della leggibilità di testi giuridici. Il caso più significativo è quello degli esperimenti condotti congiuntamente dai ricercatori IDG–ITTIG e dai ricercatori dell'Istituto di Linguistica Computazionale del CNR di Pisa, finalizzati allo sviluppo di "un software completo ed articolato, che permetta il controllo della correttezza, leggibilità e coerenza linguistica di un testo giuridico" (Biagioli et al., 1988a, p. 24). L'obiettivo era quello di "stabilire dei paradigmi di comportamento linguistico", quali la distribuzione del lessico nel testo, in grado di "offrire una misurazione globale della complessità sintattico–semantica di un testo giuridico" (Biagioli et al., 1988b, p. 49), superando in questo modo i limiti delle formule di leggibilità del testo (come la formula Flesch) sin a quel momento ampiamente utilizzate soprattutto nel contesto nord–americano del movimento 'Plain Language'.

Tale iniziativa era in particolare mossa dall'intento di offrire un'alternativa linguisticamente fondata all'idea, allora largamente diffusa negli USA, che il processo di semplificazione del corpus normativo di uno stato fosse unicamente basato sulla misura oggettiva e assoluta di un indice di leggibilità in grado di valutare anche la comprensibilità di un testo legislativo. Basati infatti sull'intuizione che la maggiore complessità di un testo sia unicamente legata alla presenza di parole e frasi lunghe, questi indici erano in realtà ampiamente discussi e criticati anche in ambito anglo–americano. È il caso, ad esempio, di Charrow et al. (1982, p. 188), in cui viene riportata la situa-

e affermazione del principio di semplificare i testi normativi, amministrativi ecc.". Per una rassegna bibliografica aggiornata dei manuali sino ad oggi redatti a livello nazionale e regionale vedi http://www.maldura.unipd.it/buro/

[19]Per una rassegna dei primi sistemi si rimanda a Biagioli (1995) e a Mercatali (2004).

zione assurda di "symplifying tax forms to an 8th–grade level, as misured by a readability formula, and then finding, as one would expect, that 8th graders cannot fill one out, or even understand it". L'obiettivo di Charrow e colleghi era infatti quello di denunciare il fatto che tale indicatore fosse fondato "in misapprehension that the number of syllables per word and the number of words per sentence are accurate indicators of the comprehensibility of a document". Al contrario, "la leggibilità è la condizione necessaria ma non sufficiente perché i testi siano chiari, semplici e precisi", come afferma Piemontese (2001, p. 128).

Tuttavia, gli strumenti di analisi automatica del testo allora a disposizione non permettevano di indagare quali fossero le caratteristiche linguistiche effettivamente rivelatrici del livello di leggibilità dei testi giuridici[20].

A fianco di questa tipologia di sistemi e strumenti legimatici finalizzati alla redazione assistita di leggi, più recentemente, come illustrato nel paragrafo seguente, sono state avviate una serie di attività complementari rivolte allo sviluppo di strumenti di controllo e gestione ex–post del testo normativo redatto. Ciò è reso possibile grazie a metodi di gestione della documentazione giuridica basati su tecniche e strumenti di Trattamento Automatico del Linguaggio.

2.3.1.1 Legimatica e Trattamento Automatico del Linguaggio

Oggi, in seguito soprattutto all'evoluzione degli strumenti di Trattamento Automatico del Linguaggio, le ricerche in questo ambito sono rivolte sempre più alla descrizione del discorso giuridico condotta a partire dall'analisi del linguaggio in cui esso è espresso in vista di successive fasi di controllo e gestione automatica del testo di legge e del suo contenuto. È l'obiettivo di ricerca perseguito da più gruppi di ricerca e chiaramente delineato da Romano (2005), il quale descrive le metodologie di ricerca condotte dai ricercatori dell'ITTIG in materia di Trattamento Automatico del Linguaggio Legislativo finalizzate alla "definizione, descrizione e formalizzazione di modelli di

[20]La metodologia di monitoraggio del profilo linguistico di testi giuridici descritta nel Capitolo 4 si inserisce appunto in questo filone di ricerche. Come discusso nel Paragrafo 4.3.2, essa mira a dimostrare come gli strumenti di annotazione linguistica automatica del testo possano essere oggi considerati come un punto di partenza affidabile per ricavare utili indicatori del grado di leggibilità di un testo (anche giuridico) a partire dalle principali caratteristiche lessicali, morfosintattiche e sintattiche in esso rintracciate.

strutture di testi normativi in base a regole linguistiche per il riconoscimento automatico e l'analisi del linguaggio dei testi normativi stessi".

Tali ricerche si basano sul modello di strutturazione suggerito (fin dalla sua prima edizione nel 1991) dal manuale di "Regole e suggerimenti per la redazione dei testi normativi". In base a quanto prescritto, un testo legislativo redatto secondo i criteri–guida suggeriti deve *i)* essere correttamente strutturato nel suo 'articolato', deve cioè essere opportunatamente suddiviso in base alle sue partizioni interne (libri, parti, titoli, capi, sezioni, articolo, commi, ecc.), e *ii)* possedere un'organizzazione organica e concettualmente omogenea del 'disposto' oggetto delle diverse disposizioni, l'insieme cioè degli elementi semantici del testo che, descrivendo il contenuto profondo delle sue funzioni regolative, ne costituiscono il profilo concettuale.

Le potenzialità espressive offerte da questo modello, articolato in un doppio livello di strutturazione (formale e semantico–funzionale) del testo legislativo, sono chiaramente descritte nei lavori di Carlo Biagioli e compiutamente esposte in (Biagioli, 2009). Tenendo separati i due livelli di organizzazione testuale, tale modello consente di definire schemi di annotazione in grado di rendere esplicite sia l'organizzazione in articoli e commi sia l'informazione semantica del testo di legge. Allo scopo di contestualizzare l'approccio all'annotazione semantico–lessicale di testi giuridici descritto nella Parte III di questo lavoro, è qui interessante ricordare i principi su cui si basa lo schema di annotazione semantica ideato da Biagioli e descritto con un ricco repertorio di esempi in (Biagioli, 2009).

È il modello denominato DAO (Disposizioni, Argomenti, Operatore) dal nome degli elementi minimi che lo definiscono. Esso si configura come un "linguaggio di descrizione" del testo legislativo finalizzato all'annotazione degli elementi che costituiscono il disposto legislativo, cioè degli "indicatori linguistici" che esprimono le disposizioni ("i frammenti dotati di senso compiuto"), gli argomenti ("le loro componenti logicamente necessarie")[21] e gli operatori (entità temporali, spaziali, logiche che "agiscono precisandolo sul significato di disposizioni e argomenti"). Esso permette di fatto lo studio della struttura logica del testo legislativo, considerato come un insieme di "frames" (le

[21]I due concetti sono esposti chiaramente da Biagioli e Pietropaoli (2003). Le Disposizioni sono descritte come "atti linguistici indipendenti che costituiscono il contenuto profondo", "gli elementi ricorrenti e rilevanti della regolazione della realtà, quali poteri, doveri, diritti, condotte, procedure, sanzioni, riparazioni, istituzioni, ecc."; gli Argomenti identificano "gli elementi ricorrenti, strutturali e rilevanti della realtà considerata: soggetti, oggetti, azioni, relazioni, processi, eventi, stati, ecc.".

Disposizioni) che contengono degli "slots", cioè "le componenti non solo più significative, ma anche logicamente necessarie". Questi ultimi sono appunto gli Argomenti di cui viene così esplicitato "il significato contingente (ruolo) in quel contesto (disposizione)". Dunque, mentre le Disposizioni sono così modellate sulla base di uno "schema di regole che cerca di esprimere i moventi ed il senso del legiferare", il contenuto degli Argomenti è rappresentato attraverso la "modellazione del lessico rilevante [...] del dominio regolato" in classi "secondo la tecnica di costruzione delle cosiddette ontologie".

Sebbene tale modello non utilizzi metodi e strumenti di riconoscimento automatico delle strutture linguistiche che esprimono Disposizioni, Argomenti e Operatori, tuttavia esso può qui essere considerato come un modello teorico di riferimento dell'approccio alla rappresentazione del significato di testi legislativi adottato in questo studio. Ne è qui inoltre condivisa la prospettiva metodologica. Come il modello DAO limitandosi ad una analisi linguistica ed "evitando una vera e propria interpretazione giuridica, [...] può svolgere il ruolo neutro di *standard* di descrizione della semantica dei testi"[22], così la metodologia rappresentazione dell'informazione semantico–lessicale contenuta in testi legislativi, descritta nel Capitolo 6 e sperimentata nel Capitolo 7, non ha nessuna velleità di interpretazione giuridica.

Da un punto di vista applicativo, il modello DAO ha avuto una parziale applicazione nello standard nazionale di annotazione (o 'marcatura') dei testi legislativi definito nell'ambito del progetto nazionale "Norme In Rete" (NIR)[23]. In seguito, sulla base di questo standard sono stati messi a punto una serie di approcci che, basati sull'uso di strumenti di Trattamento Automatico del Linguaggio, sono finalizzati all'annotazione automatica di collezioni di testi legislativi con metadati informativi relativi ad elementi sia formali (es. titolo, preambolo, sezione, articolo, comma, ecc.) sia semantico–funzionali (es. i riferimenti ad altre norme, il tipo di disposizione legislativa quale l'obbligo, la sanzione, l'eccezione, ecc.) rilevanti all'interno di un documento.

Sebbene condotti da gruppi di ricerca diversi e con metodi diversi di elaborazione automatica del linguaggio, gli approcci descritti da Bolioli et al.

[22]Biagioli (2009, p. 114).

[23]Oggi lo strumento di ricerca normativa on–line realizzato nell'ambito del progetto "Norme In Rete" è stato trasferito sul sito "Normattiva", il nuovo portale della legge vigente, attraverso cui i testi delle leggi statali, aggiornati in tempo reale, sono consultabili gratuitamente per tutti i cittadini, visitabile alla pagina http://www.normattiva.it/static/index.html

(2002), Bartolini et al. (2004), Mazzei et al. (2009) e Spinosa et al. (2009) hanno una medesima finalità in comune. L'obiettivo condiviso è infatti quello di sviluppare sistemi di annotazione dell'informazione relativa al profilo formale e semantico–funzionale di un documento normativo a partire dall'annotazione linguistica automatica del testo. La possibilità di rendere esplicita l'informazione relativa, ad esempio, a quale 'articolo' o 'comma', di un dato atto normativo, è stato 'abrogato' o 'sostituito' è realizzata grazie al riconoscimento automatico delle corrispondenti strutture linguistiche che esprimono tale informazione di modifica testuale.

Come ricordato nel Paragrafo 2.3.2.2, queste attività sono da annoverarsi tra quelle condotte dalla comunità di ricerca in materia di AI&Law e finalizzate a svolgere vari compiti di gestione dell'informazione giuridica basati sull'annotazione semantica del testo realizzata grazie all'uso di tecnologie linguistico–computazionali. Così, gli approcci seguiti da Bolioli et al. (2002) e Bartolini et al. (2004) sono indirizzati all'estrazione di informazione legislativa rilevante; mentre le metodologie messe a punto da Mazzei et al. (2009) e Spinosa et al. (2009), sebbene tra loro diverse, mirano a sviluppare un sistema di supporto al 'consolidamento' semi–automatico del testo legislativo vigente.

2.3.2 Artificial Intelligence and Law: compiti e applicazioni basati su metodi e tecniche di Trattamento Automatico del Linguaggio

Quando, agli inizi degli anni '90, iniziano a diffondersi in Italia i metodi dell'intelligenza artificiale applicata al diritto, la ricezione di tale filone di ricerca viene vista come "un passo in avanti nel processo di razionalizzazione della produzione legislativa"[24]. Tale "passo in avanti" si configura come un tentativo "di aggiungere strutture di standardizzazione dei contenuti alla normalizzazione della forma" e si concretizza nella possibilità "di fornire modelli della conoscenza come metodo per produrre modelli interpretativi". È l'inizio della diffusione degli esperimenti di costruzione dei cosiddetti sistemi esperti giuridici per lo più basati sulla conoscenza di dominio ('legal knowledge–based systems'), programmi informatici in grado di risolvere problemi con presta-

[24]Vedi Tiscornia (1995).

zioni simili a quelle di un esperto umano, esaminando un numero ampio di condizioni e costruendo dinamicamente una o più soluzioni[25].

L'approccio è quello delle attività di ricerca in AI&Law nate negli Stati Uniti tra la fine degli anni '60 e i primi anni '70. Tali attività sono inizialmente finalizzate a trovare soluzioni per realizzare compiti di 'legal reasoning' a partire dalla definizione formale di un modello di 'reasoning' basato su strutture concettuali (standardizzazioni di oggetti e eventi) e sulle strutture linguistiche che li descrivono.

La questione è posta in primo piano fin dal 1977 da McCarty (1977), nell'articolo annoverato tra i contributi fondamentali per la nascita della AI&Law. In tale articolo McCarty, descrivendo le potenzialità del programma di 'legal reasoning', TAXMAN, da lui sviluppato, afferma che la possibilità di fornire una rappresentazione concreta dei concetti giuridici che occorrono nei testi ha una stretta connessione con gli studi contemporanei in linguistica e psicologia cognitiva. Ciò è favorito dal fatto che "both disciplines, in opposition to their behaviorist predecessors, posit abstract mental structures in order to explain basic linguistic and psychological facts. In linguistics, the proposed mental structures are syntactic and semantic in nature, and they purport to organize and explain our intuitions about the grammaticality of various sentences".

Il sogno di McCarty era quello di creare un sistema esperto che utilizzando metodi di trattamento automatico del linguaggio e dell'informazione, permettesse di rendere le strutture concettuali giuridiche computabili, traducendole in programmi informatici. Tuttavia, gli strumenti di elaborazione automatica del linguaggio allora disponibili non gli consentivano di realizzare la sua idea di sviluppare "a natural language processor for corporate reorganization law".

Nonostante questo ostacolo, all'interno dell'intera comunità di ricerca era riconosciuto il fatto che un sistema di gestione automatica dell'informazione giuridica dovesse necessariamente confrontarsi con un compito di elaborazione del linguaggio naturale. Ciò è testimoniato dalle discussioni al centro del panel "AI and Legal Reasoning", organizzato durante l'IJCAI ("International Joint Conference on Artificial Intelligence") nel 1985[26], finalizzate a riflettere su come "legal reasoning is [...] heavily intertwined with natural language

[25]Per una buona sintesi delle tipologie di sistemi giuridici esperti esistenti e per una rassegna dei maggiori sistemi sino ad oggi sviluppati per lo più in Italia, vedi Lucatuorto (2006).

[26]Le discussioni tenute in occasione del panel sono riportate in (Rissland, 1985).

processing and common sense reasoning and therefore inherits all the hard problems that these imply". È significativo che lo statuto particolare del dominio giuridico venisse riconosciuto come l'ostacolo maggiore per le ricerche in intelligenza artificiale applicate al diritto, il fatto cioè che "modeling what a lawyer does is more complex than modeling experts in technical/scientific domains". In questo senso, dunque, la sfida più interessante era quella di riuscire a sviluppare un sistema giuridico esperto in grado di "represent and reason about [...] situations using a great deal of commonsense knowledge as well as technical knowledge".

Tuttavia, trent'anni più tardi, McCarty (2007), in occasione dell'ICAIL ("International Conference on Artificial Intelligence and Law") 2007, offrendo una rassegna dei tentativi di sviluppo di sistemi per la gestione e l'estrazione dell'informazione giuridica basati sulla conoscenza del dominio ('knowledge–based legal information systems'), denuncia il fatto che nessuna delle ricerche sino ad allora portate avanti "have attempted to tackle the natural language processing (NLP) problem head on, presumably because they assumed that full-scale NLP was just too difficult in a domain as complex as the law". Due anni dopo, McCarty (2009), in occasione del workshop "Natural Language Engineering of Legal Argumentation", porta l'attenzione sul fatto che è la sintassi "convoluted and unnatural" tipica del discorso giuridico a rendere vano ogni tentativo di estrazione della conoscenza condotto a partire dalla struttura linguistica sottostante l'informazione semantica.

È dunque questo il motivo principale per cui le principali attività di ricerca in AI&Law, sebbene con obiettivi applicativi diversi, si sono sino ad oggi prevalentemente concentrate sulla realizzazione di compiti di gestione dell'informazione giuridica a partire dai presupposti teorici della dottrina giuridica piuttosto che dalla concreta analisi del linguaggio usato nella formulazione dei concetti giuridici nel testo. Tale atteggiamento è testimoniato dalla tipologia di lavori presentati nelle diverse edizioni di conferenze e workshops in materia, quali DEON ("Conference on Deontic Logic in Computer Science")[27], JURIX ("Conference on Legal Knowledge and Information Systems")[28], ICAIL ("International Conference on Artificial Intelligence and Law")[29], LOAIT ("Workshop on Legal Ontologies and Artificial Intelligence Techniques")[30],

[27] http://www.defeasible.org/deon2010/previous
[28] http://www.jurix.nl/
[29] http://www.iaail.org/past-icail-conferences/index.html
[30] http://www.ittig.cnr.it/loait/

AICOL ("AI approaches to the complexity of legal systems")[31].

2.3.2.1 "NL [Natural Language] isn't the problem! NL is the object of study".

Negli ultimi anni tuttavia questo stato di cose si sta evolvendo verso una maggiore attenzione a confrontarsi con le difficoltà insite nel trattamento automatico della lingua del diritto. Wyner e van Engers (2009) espongono in modo chiaro la nuova prospettiva di ricerca: "NL [Natural Language] isn't the problem! NL is the object of study". Ciò è testimoniato dalle attività di ricerca recentemente avviate con notevole successo di interesse.

Ne è un esempio la serie di conferenze e workshops recentemente organizzati in materia di metodi e strumenti di Trattamento Automatico del Linguaggio applicati alla realizzazione di diversi compiti di gestione dell'informazione giuridica. È interessante notare come tali occasioni di incontro sul tema siano state pensate da ricercatori provenienti da entrambe le comunità di ricerca interessate, quella cioè di linguistica computazionale e quella in AI&Law.

Dal 2008 infatti è organizzato, nel contesto linguistico–computazionale del LREC ("Language Resources and Evaluation"), il workshop SPLeT ("Semantic Processing of Legal Texts")[32], finalizzato a risvegliare l'interesse per le sfide connesse con il trattamento automatico di testi giuridici rivolte soprattutto all'elaborazione automatica della conoscenza contenuta in collezioni documentali giuridiche. In particolare, il tentativo è quello di mettere in luce gli aspetti linguistico–computazionali connessi con l'elaborazione automatica del contenuto semantico dei testi giuridici. Inoltre, a partire dal JURIX 2008, si tiene il workshop NaLEA ("Natural Language Engineering of Legal Argumentation: Language, Logic, and Computation")[33], focalizzato sul dibattito inerente l'ausilio che metodi e strumenti di elaborazione automatica del linguaggio possono offrire al 'legal argumentation'. Nel 2011, in occasione della tredicesima edizione dell'ICAIL, è stato organizzato il workshop AHL-

[31] http://idt.uab.es/IVRXXIV-aicol09/

[32] Gli atti delle edizioni 2008 e 2010 dello SPLeT sono disponibili rispettivamente alle pagine: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W9_Proceedings.pdf e http://www.lrec-conf.org/proceedings/lrec2010/workshops/W23.pdf. L'edizione dello SPLeT 2009 (http://www.ittig.cnr.it/loait/program09.html) è stata organizzata congiuntamente con l'ICAIL 2009.

[33] http://nalea.org/

TL ("Applying Human Language Technology to the Law")[34] finalizzato a raccogliere i contributi di chi è impegnato ad utilizzare "HLT techniques and tools for automating knowledge extraction from legal texts and for processing legal language".

È vasta la gamma delle possibili applicazioni degli strumenti di Trattamento Automatico del Linguaggio ai vari compiti eseguiti nel campo dell'AI&Law. Tali applicazioni vanno dai tentativi semi–automatici di organizzare e modellare la conoscenza di dominio a quelli di automatizzare il processo di estrazione della conoscenza da documenti giuridici[35]. In ogni caso, le tecnologie linguistico–computazionali, sono utilizzate con l'obietivo di contribuire a superare il tradizionale collo di bottiglia che si incontra in ogni compito di acquisizione e gestione dell'informazione, quello cioè di rendere esplicita l'informazione implicitamente contenuta nella struttura linguistica di una collezione di documenti. Tali strumenti, rappresentando in modo esplicito il tessuto linguistico di un testo attraverso la sua annotazione a più livelli di analisi linguistica, forniscono infatti il punto di partenza per la realizzazione di diversi compiti di gestione dell'informazione giuridica.

Una raccolta dei più recenti contributi internazionali in materia è contenuta in (Francesconi et al., 2010) sulla scorta del quale in quanto segue sono descritti i più recenti contributi di coloro che hanno posto particolarmente in evidenza la necessità di accordare il processo di elaborazione automatica dell'informazione, nonché gli strumenti stessi di analisi linguistica automatica che vi stanno alla base, alle specificità della lingua del diritto[36].

In tali lavori è enfatizzata la disamina di quelle caratteristiche linguistiche dei testi giuridici che differenziandosi da testi di lingua comune possono costituire un ostacolo in fase di analisi linguistica e compromettere quindi la precisione delle succesive fasi di gestione automatica dell'informazione. L'obiettivo di questa rassegna è quello di evidenziare gli aspetti sino ad oggi

[34]Gli atti del workshop sono disponibili alla pagina http://wyner.info/research/Papers/AHLTL2011Papers.pdf

[35]Tradizionalmente, come ricordato da Wyner (2010), le due tipologie di compiti sono separati. Da un lato, "knowledge representation and reasoning systems, requires a knowledge base that is constructed by manual analysis [...] does not address the knowledge bottleneck, which is the extraction of information to compose the knowledge base"; dall'altro, "information extraction, addresses the bottleneck using natural language processing techniques which identify informative components of the text and annotate them". Tuttavia i due compiti possono essere complementari l'uno all'altro dal momento che "the extracted information can be represented in some knowledge base and reasoned with".

[36]La discussione che segue riprende parti di quella proposta da Venturi (2010).

tenuti in considerazione, enucleando potenzialità e limiti attuali degli strumenti di Trattamento Automatico del Linguaggio sviluppati per l'analisi di testi di lingua comune e applicati su testi giuridici.

Nell'ambito di questo tipo di attività, la metodologia pienamente esposta da Pala et al. (2010) è di particolare interesse. Gli autori dimostrano infatti come, per acquisire in modo automatico terminologia rilevante dal Codice Penale della Repubblica Ceca, sia necessario specializzare i moduli di annotazione morfologica e di lemmatizzazione della batteria di strumenti di Trattamento Automatico del Linguaggio da loro usata. Tale processo di specializzazione riguarda l'aggiornamento del lessico morfologico ad ampia copertura per la lingua ceca utilizzato dai due moduli di analisi. Esso è finalizzato al riconoscimento di termini rilevanti presenti nel Codice Penale ma prima sconosciuti. Le potenzialità di una tale specializzazione sono descritte da Pala et al. (2007), dove il modulo specializzato viene utilizzato per la costruzione di un lessico di valenza di verbi giuridici, finalizzato alla descrizione delle principali componenti di significato di tali verbi. In questo senso, l'obiettivo di Pala e colleghi è quello di dimostrare come lo studio a livello semantico di una collezione di testi giuridici possa beneficiare di una fase di specializzazione di dominio del livello di analisi morfologica.

Sempre in un'ottica di specializzazione della fase di acquisizione automatica di terminologia, si pone il contributo di Agnoloni et al. (2009) realizzato nell'ambito del progetto DALOS ("Drafting Legislation with Ontology–based Support Project")[37]. In quell'occasione, le potenzialità dello strumento di estrazione di terminologia e 'Ontology Learning' da testi (T2K, Text-to-Knowledge) (Dell'Orletta et al., 2006) sono state estese e accordate ad alcune caratteristiche delle direttive europee in lingua italiana usate come corpus di acquisizione. Tali specializzazioni hanno interessato soprattutto l'adattamento della mini–grammatica implementata in T2K per l'identificazione e l'acquisizione di unità terminologiche polirematiche. Ciò ha comportato una serie di restrizioni e/o estensioni delle regole formali parte della mini–grammatica utilizzata per l'estrazione di terminologia da corpora rappresentativi di altre varietà di lingua.

In un contesto di modellizzazione formale dell'informazione giuridica, particolare attenzione è stata dedicata da Nakamura et al. (2008) all'individuazione di particolarità sintattiche di testi giuridici in lingua giapponese. Tali specificità sono state tenute in considerazione per la definizione di specifiche

[37] http://www.dalosproject.eu/

regole di formalizzazione della struttura linguistica del testo secondo i modelli della logica formale. Tra le specificità sintattiche rintracciate, gli autori si sono focalizzati in particolare sull'analisi di un tipo specifico di nominalizzazione formata da due nomi legati dalla particella nominale *no* che indica l'esistenza di una relazione tra i due nomi. Particolarmente frequente nei testi giuridici, un tale costrutto ha richiesto un trattamento specifico per essere riconosciuto come verbo e per poter essere correttamente formalizzato come un evento.

Rispetto ai lavori sino ad ora descritti, diversa è l'attenzione dimostrata da McCarty (2007) e Walter (2009) per gli ostacoli posti dalle specificità della lingua del diritto alla fase di analisi sintattica automatica, base del successivo compito di gestione dell'informazione, un compito di annotazione semantica di decisioni giudiziarie nel primo caso, un compito di estrazione di definizioni giuridiche nel secondo caso. Mentre infatti nei lavori precedenti l'impegno alla specializzazione era indirizzato ad accordare il processo di acquisizione e gestione del contenuto informativo del testo giuridico, in questi due studi l'obiettivo è quello di identificare gli aspetti di analisi linguistica automatica più negativamente condizionati dalle caratteristiche della lingua del diritto. In entrambi i casi l'attenzione è posta pertanto sull'analisi dei tipi di errori commessi dagli strumenti di Trattamento Automatico del Linguaggio che, sviluppati per riconoscere le strutture linguistiche di testi di lingua comune, diminuiscono le loro performance di analisi quando impiegati nell'annotazione linguistica di altre tipologie di testi.

Il quesito di partenza posto da McCarty (2007) è infatti: "How accurate is Collins'parser on sentences from judicial opinions?"[38]. Dal momento che, come ricorda McCarty stesso, non esiste sino ad oggi un corpus di decisioni giudiziarie linguisticamente annotato in modo manuale da usare come riferimento, egli non è in grado di dare una risposta quantitativa alla domanda. Tuttavia, un'analisi qualitativa dei risultati dell'analisi dimostra che "applied to judicial opinions, the parser is very good on the internal structure of sentences, but it is weaker on prepositional phrase attachments and coordinated conjunctions". Tra le specificità della lingua del diritto messe in luce dagli studi linguistici in materia, sono infatti queste due delle più evidenti caratteristiche che, rendendo il discorso giuridico particolarmente complesso per l'utente umano, ne minano anche l'analisi linguistica automatica. Sulla base

[38]Il Collins'parser (Collins, 1999) è lo strumento di analisi sintattica automatica utilizzato da McCarty nel suo studio.

di questi risultati McCarty propone dunque una serie di possibili strategie da adottare in futuro per migliorare con successo l'accuratezza d'analisi del parser.

Un'analisi quantitativa dell'impatto che le peculiarità della lingua del diritto hanno sugli strumenti di Trattamento Automatico del Linguaggio è invece condotta da Walter (2009). A questo scopo, Walter ha analizzato in modo manuale una collezione di sentenze in lingua tedesca per un totale di 100 frasi, rendendone esplicita la struttura sintattica. Questo gli ha consentito di confrontare le performances del PReDS parser[39] nel riconoscimento della struttura sintattica dei testi giornalistici tedeschi, che il parser è stato addestrato ad analizzare, e di un corpus di sentenze, mettendone in luce le differenze. Il risultato del confronto ha dimostrato che la precisione di analisi diminuisce, passando da una percentuale pari all'86,74% di strutture sintattiche correttamente riconosciute a un 64%.

È sulla scia di quest'ultima tipologia di studi che si pone il Capitolo 3, finalizzato a mettere in luce gli aspetti sino ad oggi particolarmente problematici del trattamento automatico della lingua del diritto, a partire dall'accuratezza dell'annotazione linguistica di un corpus di atti normativo–amministrativi italiani condotta dagli strumenti di Trattamento Automatico del Linguaggio che rappresentano oggi lo stato dell'arte per la lingua italiana.

2.3.2.2 Trattamento Automatico del Linguaggio per l'annotazione semantica di testi giuridici

Un ruolo di primo piano nell'ambito delle attività di ricerca condotte nel campo dell'AI&Law è quello svolto dagli approcci finalizzati all'annotazione semantica di testi giuridici basata sull'annotazione linguistica automatica del testo. È questo il contesto in cui si colloca la metodologia di annotazione semantica di testi giuridici presentata nel Capitolo 7.

La centralità di questo tipo di attività è legata ai vantaggi insiti nel rendere esplicito all'interno di una collezione documentale il collegamento tra realizzazione linguistica di un testo e contenuto informativo. In un contesto di dominio, in particolare, l'annotazione semantica basata su una preliminare annotazione linguistica del testo si configura come il processo 'ponte' grazie al quale l'**informazione semantica di dominio** (la conoscenza di dominio) viene messa in collegamento con la realizzazione linguistica. È infatti questo

[39]Il PReDS parser (Braun, 2003) è lo strumento di analisi sintattica automatica utilizzato da Walter nel suo studio.

l'approccio generale che accomuna i diversi metodi sino ad oggi messi a punto per l'annotazione semantica di testi giuridici. Sebbene infatti essi siano finalizzati a svolgere compiti diversi di trattamento automatico dell'informazione giuridica, tuttavia una comune metodologia li lega. Metodologia che consiste, in una prima fase, nel rendere esplicita la struttura linguistica del testo relativa a più livelli di descrizione linguistica e, in una seconda fase, nell'associare alla realizzazione linguistica l'informazione semantica corrispondente.

I lavori più recenti condotti in questa direzione sono finalizzati a mettere a punto strategie di annotazione semantica di testi giuridici per:

- il riconoscimento automatico della struttura argomentiva di sentenze, come descritto da Palau e Moens (2009) e da Kuhn (2010);
- la generazione automatica di riassunti di decisioni giudiziarie, come descritto da Hachey e Grover (2006);
- l'acquisizione automatica di informazione relativa a 'fatti' e 'soggetti coinvolti' presenti in sentenze, come descritto da Wyner (2010) e Wyner e Peters (2010a);
- l'acquisizione automatica di definizioni giuridiche, come descritto da Walter (2009);
- il recupero automatico di sentenze, come descritto da Maxwell et al. (2009);
- l'acquisizione automatica di informazione relativa al profilo formale e semantico–funzionale di testi legislativi, come descritto da Bolioli et al. (2002), Bartolini et al. (2004), Mazzei et al. (2009) e Spinosa et al. (2009);
- la traduzione di atti normativi in modelli formali, come descritto da de Maat e Winkels (2011).

Kuhn (2010) nel suo studio utilizza una batteria di strumenti di Trattamento Automatico del Linguaggio per il riconoscimento di sintagmi chiave e di caratteristiche morfosintattiche specifiche espressione di zone testuali particolarmente significative all'interno di sentenze in lingua tedesca. Il livello di annotazione linguistica costituisce in questo modo il punto di partenza per

individuare la realizzazione linguistica di aree del testo generali e più specifiche, quali l'intestazione, la descrizione dello svolgimento del processo, i motivi della decisione, i soggetti coinvolti (es. le parti attrici, il giudice), ecc..., delle quali rendere esplicito, tramite il processo di annotazione semantica, il ruolo svolto nel dipanarsi del discorso giudiziario.

Nonostante siano tenuti in considerazione indicatori linguistici diversi, lo studio di Palau e Moens (2009) ha un obiettivo simile a quello di Kuhn (2010). La fase di annotazione sintattica automatica costituisce infatti il punto di partenza per la ricostruzione della struttura argomentativa di un corpus multilingue di sentenze. Come suggerito, è questo il primo passo verso l'acquisizione automatica di aree dell'argomentazione giudiziaria rilevanti per il giudice a fini decisionali.

In modo simile, Hachey e Grover (2006) fanno affidamento sulla fase di annotazione sintattica automatica del testo per individuare, in un corpus di decisioni giudiziarie in lingua inglese, la realizzazione linguistica specifica di aree testuali rilevanti che dovranno essere successivamente annotate con metadati semantico–informativi relativi alla funzione svolta nella sentenza da ogni specifica area (preambolo, motivazioni, ecc...). L'annotazione semantica così condotta costituisce la base per generare in modo automatico riassunti della collezione documentale di partenza.

Sempre facendo riferimento alla struttura di sentenze in lingua inglese, Wyner (2010) e Wyner e Peters (2010a) hanno messo a punto una metodologia finalizzata ad individuare nel testo i 'fatti' e i 'soggetti coinvolti' rilevanti e ad estrarli in modo automatico sulla base della loro realizzazione linguistica resa esplicita dagli strumenti di annotazione linguistica automatica.

Il lavoro di valutazione quantitativa del grado di precisione d'analisi sintattica automatica condotto da Walter (2009) è finalizzato ad individuare strutture linguistiche espressione di definizioni di concetti fondamentali presenti in sentenze in lingua tedesca. La finalità ultima è quella di rintracciare ed acquisire in modo automatico le definizioni.

Il problema della precisione degli strumenti di Trattamento Automatico del Linguaggio addestrati al riconoscimento di strutture caratteristiche della lingua comune è sollevato anche da Maxwell et al. (2009), che dimostrano che le prestazioni degli strumenti allo stato dell'arte per la lingua inglese utilizzati per l'annotazione semantica di strutture predicato–argomenti (con i relativi ruoli semantici) diminuiscono nel riconscimento di eventi presenti in sentenze. Nonostante gli autori non forniscano una valutazione nè quantitativa nè qualitativa di questa diminuzione di precisione d'analisi, le cause sono

ricondotte alle peculiarità della lingua del diritto che, differenziandosi dalla lingua comune, influiscono negativamente sui risultati della fase di recupero automatico delle sentenze basata sull'annotazione semantica.

Come precedentemente discusso nel Paragrafo 2.3.1.1, gli approcci messi a punto da Bolioli et al. (2002), Bartolini et al. (2004), Mazzei et al. (2009) e Spinosa et al. (2009) fanno affidamento su tipologie diverse di strumenti di annotazione linguistica automatica del testo per rendere esplicita la struttura linguistica di elementi formali e semantico–funzionali di atti legislativi italiani. Sebbene in maniera differente, tali approcci sono finalizzati in ultima istanza all'acquisizione automatica del contenuto informativo (legislativo) rilevante.

In modo simile, nel loro studio de Maat e Winkels (2011) hanno in un primo momento reso esplicita la struttura formale di un corpus di atti normativi olandesi, marcando le parti dell'articolato, e in un secondo momento ne hanno semanticamente annotato le frasi espressione di specifici tipi di disposizione (ad esempio, gli obblighi). A questo scopo hanno utilizzato uno strumento di annotazione sintattica a dipendenze sviluppato per l'olandese[40]. Le relazioni di dipendenza sintattica individuate in modo automatico dal parser (come ad esempio 'soggetto', 'oggetto', ecc...) sono servite come punto di partenza per annotare il ruolo semantico (ad esempio, 'agente', 'paziente', ecc...) svolto dai principali elementi informativi presenti in una frase. Una tale strategia ha permesso di organizzare l'informazione semantica contenuta in una frase, traducendo la sua realizzazione linguistica in una struttura formale.

Infine, un caso particolare di annotazione semantica basata su strumenti di Trattamento Automatico del Linguaggio è quello costituito dai lavori di Rathert (2006), Mustafaraj et al. (2006) e Wyner e Peters (2010b) descritti nel Paragrafo 6.4.3. Essi, infatti, a partire da una fase di annotazione linguistica automatica del testo, utilizzano i principali modelli di rappresentazione del significato (passati in rassegna nel Capitolo 6 di questo lavoro) per rendere esplicita l'informazione semantico–lessicale contenuta in corpora di testi giuridici.

[40]Lo strumento usato da de Maat e Winkels (2011) è l'Alpino parser (Bouma et al., 2000).

Parte II

L'annotazione sintattica di testi giuridici

Capitolo 3

Il trattamento automatico della lingua del diritto

Come messo in evidenza sin dal 1986 nella prefazione di Grishman e Kittredge (1986, p. XV), è ben noto che "except for highly circumscribed sublanguages such as weather report, we are currently not able to obtain correct sentence analyses with high reliability. [...] Reaching this goal is important if we are to develop useful applications involving more complex sublanguages".

La questione è al centro di questo capitolo, dedicato all'esame dei principali aspetti di elaborazione automatica del testo connessi con l'annotazione linguistica automatica di **testi normativi italiani**.

Come ricordato da Walter (2009), non è detto che strumenti di annotazione linguistica automatica sviluppato per il trattamento della lingua comune abbiano la stessa precisione di analisi quando utilizzati per l'elaborazione di sentenze giudiziarie. Avranno influenza infatti le caratteristiche della lingua del diritto, le quali, differenziandosi da quelle della lingua comune, devono essere trattate in modo specifico per non influire negativamente sull'accuratezza dei risultati dell'analisi.

Il tema si inserisce nel più generale interesse per gli ostacoli e le sfide posti dall'utilizzo di strumenti di Trattamento Automatico del Linguaggio per l'annotazione linguistica automatica di corpora testuali di dominio. A partire dai primi studi condotti negli anni '80, l'attenzione si è focalizzata sulla necessità di accordare le varie fasi di elaborazione automatica del testo alle specificità di un determinato linguaggio specialistico. Per questo motivo sono state sino ad oggi messe a punto una serie di strategie finalizzate ad

adattare gli strumenti utilizzati per l'elaborazione della lingua comune al riconoscimento della struttura linguistica specifica di testi di dominio.

Prendendo le mosse dalle prime attività di ricerca realizzate in quest'ambito (presentate nel Paragrafo 3.1), l'obiettivo di questo capitolo è quello di discutere il caso particolare del dominio giuridico. Dominio sino ad oggi trascurato da questa tipologia di studi, esso è al contrario un buon banco di prova per strumenti sviluppati per il trattamento automatico della lingua comune.

In quanto segue è pertanto condotta una dettagliata indagine di quegli aspetti di trattamento automatico della lingua del diritto che, relativi all'elaborazione di costrutti specifici di questa varietà linguistica, sono responsabili di un calo del livello di precisione dell'annotazione linguistica automatica di testi normativi. Particolare interesse è dedicato all'analisi di quali aspetti influiscono di più sul grado di accuratezza dell'annotazione **sintattica** automatica. Una tale attenzione è legata alla centralità che questo livello ricopre per la successiva fase di annotazione semantica, costituendone l'imprescindibile punto di partenza.

La metodologia di indagine qui messa a punto si basa sul confronto tra il livello di precisione degli strumenti di elaborazione automatica del linguaggio nell'annotazione di testi giornalistici, assunti come rappresentativi della lingua comune, e di testi normativi. L'analisi comparativa ha preso le mosse dalla raccolta di una collezione di frasi estratte da testi normativi linguisticamente annotate in modo manuale fino al livello sintattico a dipendenze. Ciò ha permesso:

- di definire una serie di specializzazioni dei criteri da seguire per annotare in modo adeguato la struttura di alcune costruzioni specifiche di dominio (Paragrafo 3.3);
- di valutare in modo comparativo il livello di precisione dell'annotazione sintattica a dipendenze realizzata dal parser statistico usato in questo studio[1] nell'annotazione di testi giornalistici e atti normativi (Paragrafo 3.4), nonché singoli aspetti particolarmente complessi di annotazione sintattica;
- di porre le necessarie premesse volte a definire una metodologia di adattamento di strumenti di Trattamento Automatico del Linguaggio all'annotazione sintattica di testi giuridici (Paragrafo 3.5).

[1] Vedi Paragrafo 3.2

3.1 Considerazioni preliminari: l'annotazione linguistica automatica di testi di dominio

I primi studi realizzati in ambito nord–americano agli inizi degli anni '80 e finalizzati all'uso di strumenti di Trattamento Automatico del Linguaggio per l'annotazione linguistica di testi caratterizzati da linguaggi specialistici si proponevano un obiettivo applicativo ben chiaro[2]. Miravano a individuare una metodologia di elaborazione automatica del testo messa a punto sulla scorta di un preliminare studio delle peculiarità di un determinato linguaggio specialistico, studio realizzato a partire dalle differenze linguistiche esistenti rispetto alle caratteristiche proprie della lingua comune.

Tale approccio aveva il suo fondamento nella "Theory of Sublanguages" di Zellig Harris[3]. Secondo Harris ogni linguaggio specialistico è descrivibile come un 'sottoinsieme' dell'insieme rappresentato dal linguaggio naturale. In analogia con la teoria matematica degli insiemi, esso può dunque essere 'operativamente' denominato **sublanguage**, inteso come "a subsystem of language that behaves essentially like the whole language, while being limited in reference to a specific subject domain" (Grishman e Kittredge, 1986, p. ix). Ogni linguaggio specialistico è pertanto studiabile nei termini di 'restrizione' o 'ampliamento', 'intersezione' o 'deviazione' rispetto alle caratteristiche proprie del linguaggio comune.

È qui d'interesse sottolineare come lo studio delle somiglianze e differenze tra un 'sublanguage' e la lingua comune avesse un duplice intento, teorico e applicativo. Da un lato, l'analisi dei 'sublanguages' come sistemi con un comportamento 'autonomo' rispetto a quello della lingua comune ne rendeva interessante lo studio sul piano teorico in quanto "microcosms of the whole language" (Grishman e Kittredge, 1986, p. x), in grado di fornire informazioni sul linguaggio naturale stesso. Dall'altro, tale interesse era finalizzato a indagare come le caratteristiche proprie di un 'sublanguage', cioè ciò in cui esso si differenzia dalla lingua comune, potessero ripercuotersi sull'accuratezza dell'annotazione linguistica automatica di testi di dominio.

In questa prospettiva lo studio delle peculiarità sintattiche riveste un interesse particolare. Come ricordato da Bonzi (1990, p. 121), infatti, "unco-

[2]Vedi tra gli altri, in particolare, gli studi di Kittredge (1982), Grishman et al. (1984), Grishman e Kittredge (1986) e Lehrberger (1986).

[3]Per una descrizione completa della teoria di Harris si rimanda a Harris (1968).

vering the regularities and major differences in the use of syntactic patterns among various disciplines and text types may help to build more efficient parsers for natural language input, to uncover better ways of automatically finding terms which best describe a document, or to better represent a user's natural language problem statement".

Nell'ambito di questi primi studi, tuttavia, i domini specialistici oggetto di maggiore interesse erano quelli caratterizzati da un linguaggio altamente specialistico come quello biomedico. È infatti questo il dominio di conoscenza al centro delle principali attività di ricerca condotte in quegli anni. Così il progetto considerato pioniere in questo ambito, il "Linguistic String Project"[4], avviato nel 1965, era espressamente finalizzato a definire una strategia di annotazione sintattica di testi di linguaggio biomedico come primo passo di un processo completo di trattamento automatico dell'informazione contenuta in testi di letteratura scientifica (Sager et al., 1987).

Ne deriva in questo modo, come notava già allora Kittredge (1982, pp. 110 e sgg.), una visione 'parziale' dell'universo dei 'sublanguages', ristretta a quelli altamente tecnico–specialistici. Ne sono invece esclusi quei domini caratterizzati da linguaggi non nettamente separati da quello comune, per i quali cioè la definizione proposta da Harris di 'sublanguage' come 'sottoinsieme' di un sistema linguistico più ampio è una condizione necessaria ma non sufficiente per stabilire operativamente i confini che separano un 'sublanguage' dalla lingua comune.

Il caso dell'esclusione della lingua del diritto da questi primi studi è un esempio significativo. È indicativo il fatto che in Kittredge (1982), dove sono raccolti i risultati delle principali attività di ricerca condotte in materia di trattamento automatico di linguaggi specialistici, un unico contributo sia dedicato al dominio giuridico. Si tratta di quello di Charrow e colleghi che in Charrow et al. (1982) focalizzano l'attenzione delle loro analisi *i)* sui fattori storici, sociologici, politici, ecc... che hanno determinato la differenza tra il "legal sublanguage" e la lingua comune e *ii)* sulle possibilità di rendere in futuro tale linguaggio più accessibile ai non addetti ai lavori[5]. A differenza dunque della maggior parte degli altri contributi dedicati a riportare i risultati di studi finalizzati a rendere più accurati strumenti di

[4]http://cs.nyu.edu/cs/projects/lsp/index.html

[5]Come ricordato nel Paragrafo 2.3.1, in Charrow et al. (1982) obiettivo polemico sono gli indici di leggibilità allora in uso che, basati unicamente su caratteristiche linguistiche superficiali del testo, non permettono di fatto di valutare l'effettivo livello di comprensibilità i un testo giuridico.

trattamento automatico del linguaggio biomedico, l'attenzione non è rivolta verso la definizione di una metodologia innovativa di analisi automatica di testi rappresentativi della lingua del diritto.

Oggi, invece, come precedentemente discusso nel Paragrafo 2.3.2.1, sempre più si sta diffondendo l'attenzione per uno studio delle potenzialità e dei limiti degli strumenti di Trattamento Automatico del Linguagggio nell'annotazione linguistica di testi giuridici. Ciò è in concomitanza con il crescente interesse per l'utilizzo di corpora sintatticamente annotati in modo automatico come punto di partenza per compiti complessi di gestione del contenuto informativo di testi giuridici.

A differenza di quanto avveniva al tempo dei primi studi volti alla definizione di strategie di adattamento degli strumenti di trattamento automatico del linguaggio per l'eborazione di linguaggi specialistici, oggi le ricerche in generale in materia di elaborazione automatica del linguaggio naturale sono basate su un diverso paradigma di annotazione sintattica. Mentre gli strumenti allora maggiormente diffusi seguivano un **grammar–driven approach**, quelli oggi usati nella gran parte delle attività di ricerca adottano un **data–driven approach**. Secondo la definizione offerta da Nivre (2006), il primo "depends on a more or less satisfactory language approximation", approssimazione definita in modo deduttivo dal linguista sulla base delle sue intuizioni, il secondo "depends on inductive inference from a more or less representative language sample" (Nivre, 2006, p. 30).

Ai fini di quanto qui discusso, è d'interesse far notare che i due diversi approcci all'elaborazione sintattica implicano un diverso comportamento degli strumenti di annotazione nel caso in cui il testo da analizzare sia scritto in un 'sublanguage' che differisce per alcune caratteristiche specifiche dalla lingua comune per il trattamento della quale gli strumenti sono stati costruiti. Sono in gioco modi diversi di soddisfare due dei requisiti fondamentali che deve avere uno strumento di annotazione sintattica ('syntactic parser') del testo: quello *i)* di essere 'robusto'[6] nel trattare input mal formato o diverso dal linguaggio per il trattamento del quale è stato sviluppato e *ii)* di essere in grado di 'disambiguare'[7] tra possibili analisi diverse. Una maggiore

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

[7]Il requisito di 'disambiguazione' ("disambiguation") di uno strumento di parsing sintattico è così definito: "A system P for parsing texts in language L satisfies the requirement

o minore robustezza e/o capacità di disambiguazione delle analisi fornite ha infatti ripercussioni sul grado di 'accuratezza'[8] dell'annotazione sintattica.

Nel primo caso, l'annotazione sintattica del testo è condotta usando grammatiche formali (come ad esempio le context–free grammars) in grado di definire un linguaggio L. L'obiettivo è quello di riconoscere l'appartenenza di ogni frase x di un testo alla grammatica definita a priori e di renderne esplicita la corretta struttura sintattica (di 'annotarne' cioè la struttura sintattica).

Nel secondo caso, l'annotazione sintattica consiste nel generare l'analisi corretta per una frase x del testo analizzato attraverso un processo di inferenza induttiva a partire da un cosiddetto 'gold standard' di riferimento, "i.e., a reference corpus of texts, where each relevant text segment has been assigned its correct analysis by a human expert" (Nivre, 2006, p. 18). È centrale pertanto per questa famiglia di strumenti d'analisi la presenza di un 'training' corpus (detto 'gold' corpus perché corretto in quanto annotato in modo manuale) dal quale gli strumenti 'apprendono' ad associare al testo l'informazione linguistica corretta grazie ad un costante processo inferenziale di classificazione probabilistica, durante il quale, ad ogni passo della computazione, viene scelta l'annotazione sintattica più probabile data la parola in input. Dal 'training' corpus gli strumenti, utilizzando algoritmi di apprendimento automatico, ricavano in questo modo un modello matematico probabilistico da applicare all'annotazione linguistica di un corpus 'sconosciuto'.

Sebbene semplificando molto, le differenze tra i due approcci rispetto ai requisiti di analisi da soddisfare sono così riassunte da Nivre (2006, p. 30): "if the grammar–based approach is sometimes characterized as being strong with respect to accuracy, but weaker with respect to robustness, disambiguation and efficiency, the reverse is often said to be true for the data–driven approach".

Nivre ricorda cioè che anche a discapito della correttezza d'analisi, l'approccio 'data–driven' consente di assegnare sempre comunque almeno un'analisi alla frase in input. Questo fa sì che anche testi caratterizzati da un linguaggio differente (come un 'sublanguage') da quello del 'training' corpus

of disambiguation if and only if, for any text T = (x_1, ..., x_n) in L, P assigns at most one analysis to every text sentence $x_i \in T$" (Nivre, 2006, p. 42).

[8]La 'accuratezza' ("accuracy") di uno strumento di parsing sintattico è così definita: "A system P for parsing texts in language L satisfies the requirement of accuracy if and only if, for any text T = (x_1, ..., x_n) in L, P assigns the correct analysis to every text sentence $x_i \in T$" (Nivre, 2006, p. 42).

di riferimento siano sempre comunque annotati. In questo modo viene superato il problema della robustezza tipico degli approcci 'grammar–driven', che per riuscire ad analizzare testi di dominio richiedono un non indifferente impegno umano per estendere la grammatica formale definita per la lingua comune.

Per supplire infatti alla mancanza di 'copertura'[9] della grammatica, il suggerimento metodologico dei primi studi nord–americani basati su di un approccio 'grammar–driven' era quello di tracciare "a refined sublanguage profile stating the relative frequencies of different sentence and text structures" (Kittredge, 1982) rispetto al profilo della lingua comune. L'attenzione per le caratteristiche di un 'sublanguage' era pertanto finalizzata alla creazione di una nuova grammatica formale, costruita a partire dagli usi sintattici tipici, diversi da quelli propri della lingua comune.

La direzione di ricerca è esposta chiaramente da Grishman et al. (1984) che dichiarano l'intenzione di trovare "a discovery procedure [...] – a procedure which can determine the domain dependent information from sample texts in the sublanguage", una strategia che permettesse di "adapt a broad–coverage grammar to the syntax of a particolar sublanguage". Una tale grammatica adattata alle specificità di un 'sublanguage' avrebbe consentito di annotare i testi rappresentativi di un determinato dominio specialistico con una accuratezza di analisi maggiore di quella ottenuta utilizzando una grammatica sviluppata per trattare testi di lingua comune.

Nel caso degli approcci 'data–driven', il problema principale rimane quello di stabilire quale sia tra le varie analisi generate dal sistema per una frase x quella corretta in determinato contesto d'uso, di 'disambiguare' cioè quale sia l'unica possibile analisi corretta di x per il tipo di testo nel quale x occorre. Nella situazione ottimale, la questione è risolta addestrando gli strumenti di analisi del testo su di un 'training' corpus composto da testi rappresentativi di una determinata varietà testuale o linguistica. Dati i notevoli costi, sia in termini di tempo sia di impegno di lavoro, richiesti nella costruzione di una risorsa sintatticamente annotata in modo manuale, non sempre tuttavia un 'gold' corpus per una nuova varietà è disponibile. Questo ad oggi costituisce il principale ostacolo alle attività di adattamento di strumenti di annotazione linguistica del testo a domini specialistici caratterizzati da un linguaggio diverso da quello dei testi sui quali gli strumenti sono stati addestrati. Il caso

[9]Per 'copertura' di una grammatica si intende la sua capacità di tenere conto dei fenomeni sintattici propri del linguaggio che appartiene alla grammatica.

dell'adattamento finalizzato al trattamento automatico di 'sublanguage' è in questo senso emblematico.

Sebbene il tema dell'adattamento di strumenti di Trattamento Automatico del Linguaggio all'elaborazione di 'sublanguages' rappresenti una sfida tutt'ora aperta sia per approcci 'grammar–driven' sia per quelli 'data–driven', in questo studio ci si concentrerà sulle questioni connesse con l'annotazione linguistica automatica di testi giuridici realizzata da strumenti del secondo tipo, ovvero basati su algoritmi di apprendimento automatico da dati testuali.

Come dimostrato infatti a livello internazionale nelle più recenti edizioni della "Conference on Computational Natural Language Learning" (CoNLL)[10], nell'ambito della quale vengono organizzate campagne di valutazione di strumenti per l'analisi automatica del linguaggio naturale sviluppati per diverse lingue, è infatti questo il tipo di strumenti di Trattamento Automatico del Linguaggio che si sta diffondendo sempre di più, dimostrando le migliori prestazioni di annotazione testuale. Anche per la lingua italiana un tale paradigma di analisi si sta affermando con risultati sempre più affidabili, come testimoniano i risultati del "Parsing Track" dell'ultima edizione di "Evalita" 2009[11], la campagna di valutazione di strumenti di annotazione linguistica automatica per l'italiano.

3.2 La catena di strumenti di Trattamento Automatico del Linguaggio

In linea con lo stato dell'arte tratteggiato nel precedente paragrafo, la catena di strumenti di annotazione linguistica automatica per l'italiano utilizzata in questo lavoro segue un approccio 'data–driven' all'annotazione linguistica del testo.

Essa è composta da una serie di strumenti di Trattamento Automatico del Linguaggio sviluppati dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa, che, operando in successione, permettono di rendere progressivamente esplicita l'informazione linguistica contenuta in un testo. Per ogni livello di descrizione linguistica uno specifico componente di analisi identifica in modo automatico la strut-

[10]Vedi in particolare gli atti delle edizioni 2007, 2008 e 2009 accessibili dalla pagina http://ifarm.nl/signll/conll/

[11]http://www.evalita.it/2009/proceedings

tura del testo, utilizzando come input il risultato prodotto dal componente precedente.

Il testo viene così annotato a più livelli di analisi, rendendo incrementalmente esplicite le seguenti informazioni:

a) i singoli periodi che compongono il testo (segmentazione del testo in periodi);

b) le singole parole ortografiche ('tokens') presenti, compresi i segni di punteggiatura (tokenizzazione);

c) la categoria morfosintattica rilevante nel contesto specifico e il lemma corrispondenti ad ogni singolo token (disambiguazione morfosintattica[12] e lemmatizzazione);

d) la struttura sintattica secondo una rappresentazione a dipendenze (annotazione delle relazioni di dipendenza sintattica)[13].

Centrale per le successive discussioni contenute in questo lavoro è la fase di annotazione sintattica, che consente di descrivere la struttura di un periodo sotto forma di relazioni binarie di dipendenza tra tokens. Detto con le parole di Nivre: "The fundamental notion of dependency is based on the idea that the syntactic structure of a sentence consists of binary asymmetrical relations between the words of the sentence" (Nivre, 2006, p. 47), relazioni che sussistono tra una testa sintattica e il suo dipendente. In termini di rappresentazione computazionale delle relazioni di dipendenza, questo approccio all'analisi sintattica implica che in fase di annotazione ogni periodo sia rappresentato come una serie di "lexical nodes, connected by dependency arcs, possibly labeled with dependency types" (Nivre, 2006, p. 55).

La Tabella 3.1 mostra un esempio del risultato del processo incrementale di annotazione linguistica del periodo

- *Le disposizioni di cui alla presente lettera si applicano anche nei confronti degli altri organi tenuti all'adozione di strumenti urbanistici.*

Innanzitutto, il periodo è stato individuato grazie alla fase di segmentazione in periodi di una direttiva comunitaria in materia ambientale. Durante

[12] 'Part–of–Speech Tagging'.
[13] 'Dependency parsing'.

la successiva fase di tokenizzazione, all'interno del periodo sono stati riconosciuti i tokens corrispondenti alle singole forme (seconda colonna), identificate univocamente da un numero progressivo (prima colonna). La fase di disambiguazione morfosintattica ha permesso di associare ad ogni token individuato *i)* la corretta categoria morfosintattica (quarta e quinta colonna)[14] che il token ha nel contesto specifico, *ii)* i relativi tratti morfologici (sesta colonna) e *iii)* il lemma corrispondente (terza colonnna). Ad esempio, la forma *disposizioni* (Id=2) viene ricondotta al lemma *disposizione*, viene annotato con la categoria sostantivo (S) e viene inoltre riconosciuto che si tratta di una forma plurale (num=p) e femminile (gen=f).

		Lemmatizzazione	Annotazione morfosintattica			Annotazione sintattica	
Id	Forma	Lemma	CPoS	FPoS	Tratti	Testa	Relazione
1	Le	il	R	RD	num=p|gen=f	2	det
2	disposizioni	disposizione	S	S	num=p|gen=f	9	subj
3	di	di	E	E	_	5	comp
4	cui	cui	P	PR	num=n|gen=n	3	prep
5	alla	al	E	EA	num=s|gen=f	2	mod_rel
6	presente	presente	A	A	num=s|gen=n	7	mod
7	lettera	lettera	S	S	num=s|gen=f	5	prep
8	si	si	P	PC	num=n|per=3|gen=n	9	clit
9	applicano	applicare	V	V	num=p|per=3|mod=i|ten=p	0	ROOT
10	anche	anche	B	B	_	9	mod
11	nei	in	E	EA	num=p|gen=m	9	comp
12	confronti	confronto	S	S	num=p|gen=m	11	prep
13	degli	di	E	EA	num=p|gen=m	12	comp
14	altri	altro	A	A	num=p|gen=m	15	mod
15	organi	organo	S	S	num=p|gen=m	13	prep
16	tenuti	tenere	V	V	num=p|mod=p|gen=m	15	mod
17	all'	a	E	EA	num=s|gen=n	16	comp
18	adozione	adozione	S	S	num=s|gen=f	17	prep
19	di	di	E	E	_	18	comp
20	strumenti	strumento	S	S	num=p|gen=m	19	prep
21	urbanistici	urbanistico	A	A	num=p|gen=m	20	mod
22	.	.	F	FS	_	9	punc

Tabella 3.1: Un esempio di annotazione della catena di analisi.

Il risultato dell'annotazione sintattica riportato nella settima e ottava colonna della Tabella 3.1 permette inoltre di stabilire che, ad esempio, il sostantivo *disposizioni* è il soggetto (subj) del verbo *applicano*, il quale costituisce la testa sintattica della relazione. Questa informazione è riportata

[14]Per ogni token viene riconosciuta la categoria morfosintattica generale (CPoS) e eventuali sottocategorie (FPoS). Ad esempio, alla forma (token) *alla* viene associata la categoria preposizione (E) e viene ulteriormente specificato che si tratta di una preposizione articolata (EA). Allo stesso modo, il token . viene annotato come un segno di punteggiatura (F) di fine periodo (FS).

L'inventario completo delle categorie morfosintattiche e delle relazioni di dipendenza contenute nello schema di annotazione degli strumenti utilizzati in questo studio è riportato nell'Allegato I.

nella settima colonna dove è infatti segnalato che la testa sintattica del dipendente *disposizioni* ha Id=9, l'Id cioè del token 'applicano'. In questo caso *applicano* ha testa sintattica 0 dal momento che rappresenta il verbo della frase principale, radice (root) dell'albero sintattico dell'intero periodo.

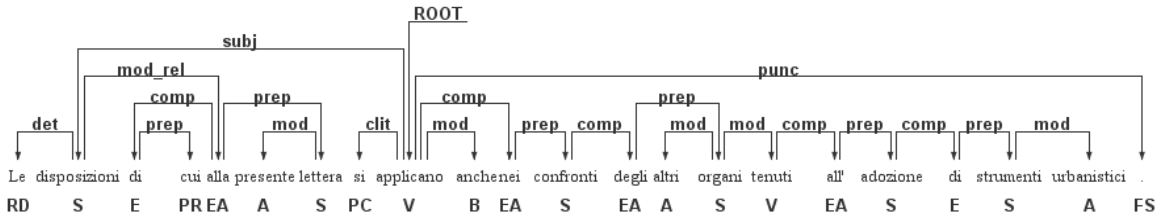


Figura 3.1: Un esempio di rappresentazione grafica dell'annotazione sintattica a dipendenze.

La fase di annotazione sintattica a dipendenze permette dunque di fornire una descrizione esplicita dell'intero albero sintattico del periodo analizzato, sotto forma di relazioni di dipendenza che legano i tokens che lo compongono. L'informazione può inoltre essere graficamente visualizzata, come mostra la Figura 3.1 che riporta la struttura sintattica della frase annotata, rappresentata come una serie di 'nodi' lessicali (i singoli tokens), messi in collegamento da 'archi' di dipendenza a loro volta etichettati con il nome del tipo di relazione di dipendenza (gli archi e le etichette graficamente rappresentati).

All'interno della catena di annotazione, ciascuna fase di analisi linguistica automatica è realizzata da un singolo componente. In particolare, l'annotazione morfosintattica è realizzata dal modulo di analisi descritto da Dell'Orletta (2009) e l'annotazione sintattica a dipendenze dal parser sintattico DeSR (Attardi et al., 2009).

In entrambi i casi si tratta di strumenti basati su algoritmi di apprendimento automatico 'supervisionato' e rappresentano lo stato dell'arte per la lingua italiana. Le analisi sono cioè realizzate sulla base del risultato di un processo di inferenza induttiva condotto a partire da un corpus di addestramento (o 'training corpus') annotato in modo manuale, dal quale gli strumenti di annotazione apprendono a riconoscere la corretta categoria morfosintattica associata ad ogni token del periodo analizzato e a ricostruire la struttura sintattica di un intero periodo.

In particolare, il parser DeSR svolge un compito, detto con le parole di Nivre, di "inductive dependency parsing", così definito "for the general idea of using inductive machine learning to predict the actions of a dependency parser" (Nivre, 2006, p. 4). Il processo di inferenza induttiva è condotto sulla base della treebank ISST–TANL[15], la porzione di 3.109 frasi (per un totale di 71.285 tokens) di testi giornalistici estratta dalla "Italian Syntactic–Semantic Treebank" (ISST) (Montemagni et al., 2003), annotata in modo manuale e usata come 'training' corpus. Di fatto, DeSR analizza sintatticamente i periodi di un testo 'sconosciuto' sulla base delle strutture sintattiche che è stato addestrato a riconoscere nel 'training' corpus.

Nell'ambito dell'edizione 2009 della campagna di valutazione Evalita, entrambi i componenti di annotazione linguistica usati in questo studio sono risultati gli strumenti più precisi e affidabili nell'analisi automatica dell'italiano.

In particolare, il modulo di annotazione morfosintattica con un'accuratezza[16] del 96,34% si è classificato primo nel "Part–Of–Speech Tagging Task" della campagna di valutazione[17], dimostrandosi in questo modo il più preciso analizzatore morfosintattico (PoS tagger) oggi esistente per la lingua italiana.

In quell'occasione, nell'ambito del "Dependency Parsing Track"[18] il componente di analisi sintattica a dipendenze (DeSR) è risultato il parser con le migliori prestazioni di analisi alla pari del Turin University Parser (TUP) sviluppato presso l'Università di Torino (Lesmo, 2009) e basato su di un approccio 'grammar–driven'. L'oscillazione tra l'88,73% di triple [testa sintattica, dipendente, relazione di dipendenza], correttamente annotate, ottenuto da TUP e l'88,67% di DeSR non costituisce infatti una variazione di risultati statisticamente significativa.

---

[15] http://medialab.di.unipi.it/wiki/SemaWiki.

Si tratta di una versione rivista della treebank ISST–CoNLL usata nello "Shared Task on Dependency Parsing, multilingual track" della "Conference on Computational Natural Language Learning" (CoNLL 2007) (Nivre et al., 2007a).

[16] L'accuratezza è calcolata come il rapporto tra il numero di tokens classificati correttamente e il numero totale di tokens analizzati.

[17] Una descrizione generale del Task di Evalita 2009 è disponibile alla pagina http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2009/PoSTagging/POS_ORGANIZERS.pdf

[18] La descrizione completa del Task è disponibile alla pagina http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2009/Parsing/Dependency/DEP_PARS_ORGANIZERS.pdf

Le prestazioni del parser DeSR, al centro dell'analisi condotta in questo capitolo, sono descritte in dettaglio nei paragrafi successivi.

3.3 L'annotazione sintattica: la creazione di un corpus di riferimento di atti normativi per la lingua italiana

In base alle precedenti discussioni, risulta chiaro che per poter quantificare l'accuratezza degli strumenti di annotazione 'data–driven' nell'analisi di testi normativi occorre avere a disposizione un 'gold' corpus di riferimento, annotato in modo manuale rispetto al quale confrontare le performances dell'annotazione automatica. Dal momento che, tuttavia, sino ad oggi non esiste per la lingua italiana un corpus di questo tipo, è stato necessario costruirne uno.

L'unica eccezione è rappresentata dalla porzione della Turin University Treebank (TUT)[19], costruita presso l'Università di Torino, costituita da articoli del Codice Civile Italiano (per un totale di 1.100 frasi e 28.048 tokens), annotata con informazione sintattica a dipendenze in modo manuale.

Ai fini dello studio qui condotto si è deciso tuttavia di non prendere la TUT come riferimento per la valutazione dell'accuratezza di DeSR su testi del dominio giuridico dal momento che, come dimostrato da Garavelli (2001), il Codice Civile presenta numerose caratteristiche grammaticali diverse da quelle della tipologia di atti normativi qui in esame. In particolare, leggi, decreti, regolamenti oggetto delle analisi di questo lavoro si contraddistinguono, a differenza del Codice Civile, per una minore coerenza "nell'osservanza dei vincoli formali [...], ma più latamente in tutti i settori dell'organizzazione sintattica e testuale, mentre si scoprono più evidenti i segni dell'assuefazione agli stereotipi di un'ufficialità che ripiega sugli pseudospecialismi" (Garavelli, 2001, pp. 85–86).

Dunque, le caratteristiche linguistiche di leggi, decreti, regolamenti differendo maggiormente da quelle della lingua comune costituiscono una sfida maggiore per strumenti di Trattamento Automatico del Linguaggio addestrati su testi giornalistici. È questo il motivo per cui si è scelto di costruire un corpus di riferimento composto di testi giuridici di questo tipo. Il corpus,

[19] http://www.di.unito.it/~tutreeb/

d'ora in avanti chiamato 'AMBnorm–gold'[20], è una collezione di 148 frasi (per un totale di 5.691 tokens) estratte dal corpus di atti normativi e amministrativi in materia ambientale emessi da tre diverse autorità, descritto nel capitolo successivo al Paragrafo 4.1.2[21], annotate in modo automatico con DeSR e riviste in modo manuale[22].

La scelta di partire nella costruzione di AMBnorm–gold da una prima fase di annotazione automatica è motivata dal vantaggio di fare affidamento sulle annotazioni automatiche per ridurre i margini di arbitrarietà dell'annotazione manuale. In questo modo è stato possibile mantenere uniforme e coerente la revisione manuale delle annotazioni precedentemente realizzate in modo automatico. Inoltre, ciò ha permesso di individuare gli aspetti di maggiore ostacolo al trattamento automatico della lingua del diritto e, di conseguenza, di mettere in luce le specializzazioni necessarie.

Grazie alla creazione di AMBnorm–gold è stato così possibile confrontare il grado di accuratezza di DeSR nell'annotazione di testi giornalistici, sui quali il parser è stato addestrato, e di atti normativi, rappresentativi di una varietà linguistica 'sconosciuta' per il parser. Nel primo caso, come 'test' corpus è stata utilizzata una porzione di 231 frasi, per un totale di 5.166 tokens, della treebank ISST–TANL (d'ora in avanti ISST–TANL–test)[23]. Nel secondo caso, è stato usato l'output del parser nell'annotazione della collezione di frasi giuridiche in materia ambientale.

In quanto segue sono pertanto descritti gli usi che sono stati fatti di AMBnorm–gold nell'ambito di questo studio. Da un lato, il corpus è servito per individuare le specializzazioni dei criteri di annotazione necessarie per personalizzare l'intero processo di analisi linguistica al trattamento di testi normativi (Paragrafo 3.3.1). Dall'altro, il corpus ha permesso di valutare

[20]Nota che l'uso di 'gold' è in linea con la terminologia comunemente utilizzata per riferirsi alle annotazioni realizzate in modo manuale, che si assume siano per questo motivo sempre corrette.

[21]AMBnorm–gold è composto in particolare da 92 frasi del corpus 'AMBnorm(Stato)', 31 frasi del corpus 'AMBnorm(Europa)' e 25 frasi del corpus 'AMBnorm(Regione)'.

[22]La revisione manuale dell'annotazione automatica è stata condotta grazie al "Dependency Grammar Annotator" (DgAnnotator), uno strumento di visualizzazione grafica per l'annotazione di strutture sintattiche a dipendenze, messo a punto nell'ambito delle attività di sviluppo del parser DeSR. DgAnnotator è liberamente disponibile alla pagina http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/

[23]Si tratta della porzione utilizzata come 'test' in Evalita 2009 nell'ambito del "Dependency Parsing Track".

l'accuratezza degli strumenti nell'annotazione sintattica dei testi normativi, mettendone in luce gli aspetti più problematici (Paragrafo 3.4).

3.3.1 Le specializzazioni dei criteri di annotazione

Le specializzazioni dei criteri di annotazione ha preso le mosse dalla necessità di trattare alcune strutture specifiche della lingua del diritto che, contenute in AMBnorm–gold, non erano coperte dai criteri attuali adotatti nella treebank ISST–TANL.

Si tratta di aspetti che riguardano i vari livelli di annotazione linguistica del testo. Sono infatti coinvolti *i)* aspetti di segmentazione del testo in periodi legati ad una particolare organizzazione testuale del documento giuridico, *ii)* usi di parole in contesti specifici di dominio diversi da quelli propri della lingua comune, *iii)* costruzioni sintattiche particolari.

3.3.1.1 La segmentazione del testo in periodi

La revisione dei criteri di segmentazione del testo in periodi ha tenuto conto di una ben nota peculiarità dei parsers a dipendenze basati su algoritmi di apprendimento automatico. Come empiricamente dimostrato da McDonald e Nivre (2007), il tipo di algoritmo di annotazione sintattica con il quale è stato sviluppato DeSR ha una significativa diminuzione di prestazioni nel riconoscimento di periodi molto lunghi.

Ciò è principalmente dovuto, secondo McDonald e Nivre, alla lunghezza delle relazioni di dipendenza che legano un token$_i$ a un token$_j$, cioè alla distanza misurabile in numero di tokens tra un dipendente e la sua testa sintattica. Periodi molto lunghi (in termini di tokens) sono tipicamente caratterizzati, ad esempio, da dipendenti molto distanti dalla radice verbale dell'albero sintattico del periodo (root). Questo, aumentando il numero di possibili scelte di annotazione, genera ambiguità di analisi che si ripercuotono negativamente sulla precisione del processo di annotazione sintattica.

Nel caso dei testi normativi, oltre agli esempi menzionati sopra, è la presenza di lunghe catene di strutture coordinate a rappresentare una delle cause principali di ambiguità nell'analisi sintattica. Sebbene infatti la coordinazione costituisca in generale un caso piuttosto complesso da trattare in modo automatico, tuttavia la situazione è resa ancora più difficile nei testi giuridici dalla presenza di lunghi elenchi tra loro coordinati e tutti legati ad un unico elemento testa. Un esempio è rappresentato dal seguente periodo:

- *Al verificarsi di un incidente rilevante, il gestore è tenuto a:*

  a) adattare le misure previste dal piano di emergenza di cui all'articolo 11;

  b) informare il prefetto, il sindaco, il comando provinciale dei Vigili del fuoco il presidente della giunta regionale e il presidente dell'amministrazione provinciale comunicando, non appena ne venga a conoscenza:

    1) le circostanze dell'incidente;

    2) le sostanze pericolose presenti;

    3) i dati disponibili per valutare le conseguenze dell'incidente per l'uomo e per l'ambiente;

    4) le misure di emergenza adottate;

    5) le informazioni sulle misure previste per limitare gli effetti dell'incidente a medio e lungo termine ed abitare che esso si riproduca;

  c) aggiornare le informazioni fornite, qualora da indagini più approfondite emergessero nuovi elementi che modificano le precedenti informazioni o le conclusioni tratte.

In questo caso, sulla base dei criteri di segmentazione del testo in periodi seguiti in ISST–TANL, l'intero elenco è considerato un unico periodo. In questo modo i diversi elementi, ad esempio, dell'elenco alfabetico (*a), b), c)*) sono considerati frasi argomentali tra loro coordinate. Sulla base dello schema di annotazione adottato, dunque, *adattare* è il token dipendente della testa sintattica *tenuto* (radice dell'intero periodo) rispetto alla quale svolge una funzione sintattico–funzionale 'arg'; *informare* e *aggiornare*, sono i dipendenti della testa sintattica *adattare* rispetto alla quale svolgono una funzione 'conj'; i tokens di punteggiatura *;* (che segnalano graficamente la struttura dell'elenco) sono i dipendenti di *adattare* rispetto al svolgono un funzione 'con'.

Una tale scansione in elenchi fa tuttavia sì che il parser sia obbligato a ricostruire delle relazioni di dipendenza molto lunghe. Ad esempio, la relazione che lega il token coordinato *aggiornare* con la sua corrispondente testa sintattica *adattare* avrebbe una distanza di 100 tokens.

Considerata l'estrema difficoltà per il parser di riconoscere strutture coordinate tali, si è scelto di segmentare periodi di questo tipo in periodi più brevi,

costituiti dai singoli elementi coordinati parte degli elenchi, lasciando l'annotazione delle relazioni tra periodi ad un seconda fase, al momento non ancora realizzata ma che potrà esserlo in futuro.

In questo caso dunque, ad esempio, i tre elementi dell'elenco precedente (*a) adattare ...;*, *b) informare ...;* e *c) aggiornare ...* .) sono stati considerati tre periodi annotati separatamente. E, di conseguenza, *Al verificarsi di un incidente rilevante, il gestore è tenuto a:* è considerato a sua volta come un periodo a sé stante.

Questo permette di conservare l'originale struttura grafica del testo giuridico e la corrispondente organizzazione dell'informazione in esso contenuta[24]. Ogni periodo parte dell'elenco alfabetico contiene infatti un **dovere** che *il gestore* è obbligato ad adempiere; così come ogni periodo dell'elenco numerico veicola il tipo di informazione che egli è tenuto a comunicare *al prefetto, al sindaco, al comando provinciale dei Vigili del fuoco, al presidente della giunta regionale e al presidente dell'amministrazione provinciale.*

3.3.1.2 La specializzazione del lessico

Tenuta in considerazione la tendenza propria dei testi giuridici di fare un uso specialistico di termini della lingua comune è stato necessario estendere il lessico utilizzato in fase di disambiguazione morfosintattica e lemmatizzazione. A partire cioè dagli errori commessi nella fase preliminare di annotazione automatica, sono stati individuati i termini e i loro contesti di dominio che gli strumenti non erano stati addestrati a riconoscere poiché ricorrono con frequenza minore (o non ricorrono affatto) in ISST–TANL .

È il caso ad esempio del termine *direttiva*, che sistematicamente usato in ISST–TANL come aggettivo non era stato riconosciuto come sostantivo. O è anche il caso di *data* e *allegato*, per i quali è stato necessario estendere il lessico, dal momento che occorrendo più frequentemente in contesti comuni come participi passati rispettivamente del lemma *dare* e *allegare* non erano stati riconosciuti come sostantivi.

L'importanza di questo tipo di specializzazione è tanto più evidente se si considera che il risultato dell'annotazione morfosintattica costituisce l'input del successivo livello, gli errori a questo livello di analisi rischiano di inficiare il processo di annotazione sintattica.

[24]Per le conseguenza che questo comporta in fase di annotazione semantica vedi il Paragrafo 7.3.2.

3.3.1.3 L'annotazione delle relazioni di dipendenza sintattica

In questo caso, le specializzazioni dei criteri di annotazione riguardano i tipi di costruzioni per i quali non sono state sistematicamente generate le analisi corrette durante la fase preliminare di annotazione automatica. Come si può vedere dagli esempi che seguono, si tratta per lo più di strutture sintattiche tipiche dei testi normativi scarsamente rappresentate (o non rappresentate affatto) in ISST–TANL.

i) **Costruzioni ellittiche,** spesso usate nei rimandi espliciti ad altri atti normativi o a parti dell'articolato. Un esempio è rappresentato dal seguente periodo:

- *I decreti legislativi di cui al comma 1 si conformano, nel rispetto dei principi e delle norme comunitarie e delle competenze per materia delle amministrazioni statali, nonché delle attribuzioni delle regioni e degli enti locali, come definite ai sensi dell'articolo 117 della Costituzione, della legge 15 marzo 1997, n. 59, e del decreto legislativo 31 marzo 1998, n. 112, e fatte salve le norme statutarie e le relative norme di attuazione delle regioni a statuto speciale e delle province autonome di Trento e di Bolzano, e del principio di sussidiariet; ai seguenti principi e criteri direttivi generali.*

L'attenzione è posta sulla frase relativa *di cui al comma 1* con ellissi del verbo, che in AMBnorm–gold è annotata come mostrato nella Figura 3.2.

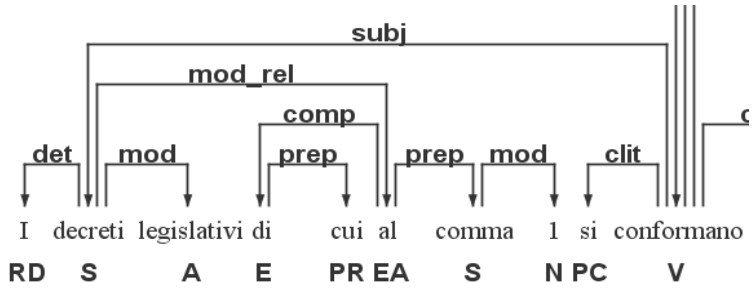


Figura 3.2: Un esempio di costruzione ellittica che coinvolge un rimando esplicito a parte dell'articolato.

In questo caso è stato necessario supplire alla mancanza del verbo nella frase relativa, il quale, sulla base dello schema di annotazione di ISST–TANL, sarebbe stato il token dipendente dall'antecedente relativo *decreti*, ad esso legato da una relazione di dipendenza 'mod_rel'. La relazione è stata allora riconosciuta tra la testa sintattica *decreti* e il token *al*, a sua volta testa del sintagma preposizionale *al comma 1.*

Un altro tipo di costruzione ellittica è quello rappresentato da occorrenze di frasi participiali ellittiche, come nel seguente esempio:

- *A norma dell'articolo 4, paragrafo 1, lettera a), punto ii), della direttiva 2000/60/CE gli Stati membri sono tenuti a proteggere, migliorare e ripristinare tutti i corpi idrici superficiali al fine di raggiungere un buono stato delle acque superficiali entro 15 anni dall'entrata in vigore della direttiva, salve alcune eccezioni, in base alle disposizioni dell'allegato V della medesima.*

La frase participiale che ha richiesto di essere trattata in modo particolare è *salve alcune eccezioni*, annotata in AMBnorm–gold come mostrato nella Figura 3.3.

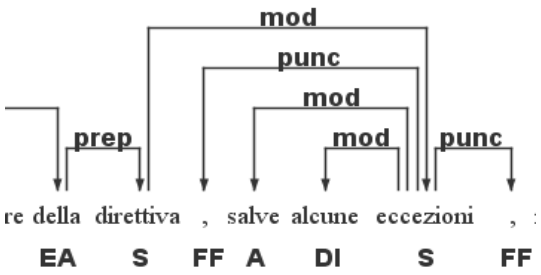


Figura 3.3: Un esempio di frase participiale ellittica.

In mancanza della forma participiale *fatte* nella frase, è stata riconosciuta una relazione 'mod' che lega il token *eccezioni* alla testa sintattica *direttiva.*

***ii)* Frasi participiali,** spesso usate in funzione eccettuativa o limitativa, come esemplificato nel periodo seguente, dove la frase *fatti salvi ogni altro adempimento o comminatoria previsti dalle leggi vigenti* è stata annotata in AMBnorm–gold come illustrato nella Figura 3.4:

- *Nel caso di violazione del disposto del comma 1, l'amministrazione competente dispone la cessazione dell'utenza abusiva ed il contravventore, fatti salvi ogni altro adempimento o comminatoria previsti dalle leggi vigenti, è tenuto al pagamento di una sanzione amministrativa pecuniaria da lire cinque milioni a lire cinquanta milioni.*

In questo caso è stata annotata una relazione 'mod' tra la testa sintattica della frase participiale (*fatti*) e il token *tenuto*. È stata così resa esplicita la relazione di dipendenza della frase participiale dalla frase *il contravventore ... è tenuto al pagamento di una sanzione amministrativa pecuniaria da lire cinque milioni a lire cinquanta milioni* di cui *tenuto* è testa.

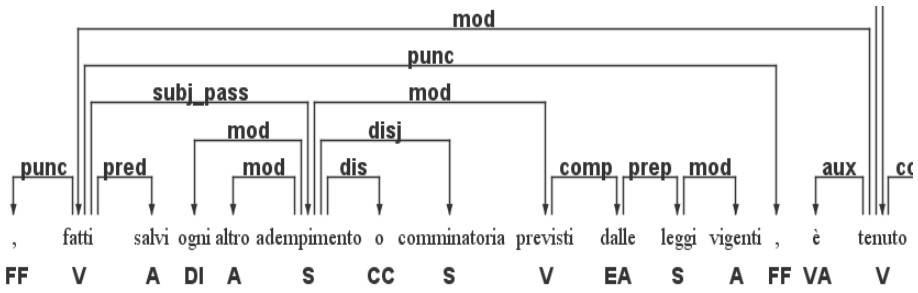


Figura 3.4: Un esempio di frase participiale eccettuativa.

***iii)* Partizioni interne dell'articolato di un atto normativo,** che, occorrono nei testi giuridici sotto forma di elenchi i cui sintagmi sono "legati tra loro dalla relazione determinato–determinante" (Garavelli, 2001, p. 79). Per questo motivo, si è scelto di annotarle come sequenze di strutture appositive, rispettando la funzione di modificazione gerarchica 'a cascata' che ogni sottopartizione svolge rispetto alla partizione precedente. La porzione di testo *articolo 94, comma 3, lettera a) della l.r. 44/2000* del seguente periodo è pertanto annotata come riportato nella Figura 3.5:

- *L'inosservanza delle disposizioni richiamate ai commi 1, 2 e 3, oltre ad essere punite con le sanzioni amministrative previste, comportano l'obbligo del ripristino, che dovrà essere realizzato in conformità alle disposizioni formulate in apposito provvedimento della Provincia di Biella, ai sensi dell'articolo 94, comma 3, lettera a) della l.r. 44/2000.*

La relazione 'mod' riconosciuta tra *articolo* (testa sintattica) e *comma* (dipendente) e tra *comma* (testa sintattica) e *lettera* (dipendente) permette di rendere esplicita la struttura gerarchica dell'articolato della legge regionale 44/2000, ordinata in articoli, commi e lettere. Inoltre la relazione 'comp' tra il token *della* e *articolo* (testa sintattica) chiarisce quest'ultima informazione, cioè che si sta facendo riferimento all'articolato di quella specifica legge.

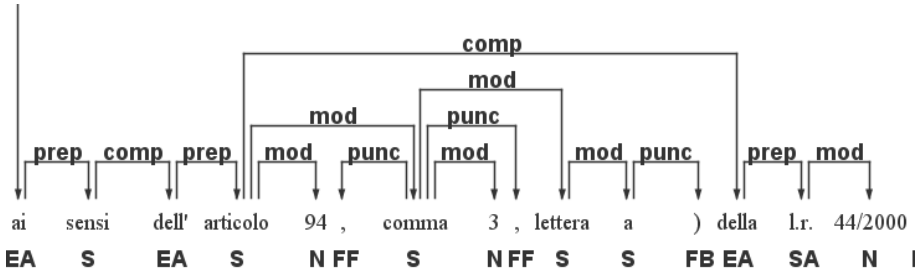


Figura 3.5: Un esempio di annotazione delle partizioni interne dell'articolato di un atto normativo.

3.4 L'analisi dell'accuratezza dell'annotazione sintattica di atti normativi

L'accuratezza di DeSR è stata testata facendo riferimento alle metriche ufficiali usate per la valutazione dei risultati di un compito di annotazione sintattica a dipendenze. Si tratta del calcolo del:

1. Labelled Attachment Score (LAS), la percentuale di tokens ai quali il parser ha assegnato correttamente sia la testa sintattica sia il tipo di relazione di dipendenza;

2. Unlabelled Attachment Score (UAS), la percentuale di tokens ai quali il parser ha assegnato in modo corretto unicamente la testa sintattica;

3. Label Accuracy score (LA), la percentuale di tokens ai quali il parser ha assegnato in modo corretto unicamente il tipo di relazione di dipendenza.

Mentre dunque il calcolo del LAS permette di valutare se il parser ha correttamente riconosciuto l'**intera** struttura sintattica di un periodo, il UAS ci dà indicazioni solo riguardo alle volte che durante l'analisi è stata correttamente riconosciuta l'esistenza di una relazione di dipendenza senza tuttavia specificarne il tipo. Il calcolo del LA è finalizzato a valutare la percentuale di relazioni di dipendenza correttamente annotate a prescindere dalla corretta individuazione della testa sintattica coinvolta. Per questo motivo, come si vedrà in quanto segue, i valori di LAS, indicativi di una valutazione più restrittiva delle performance del parser, sono sempre inferiori a quelli di UAS e LA.

Sono queste le metriche ufficiali usate in occasione dello "Shared Task on Dependency Parsing, multilingual track"[25] della "Conference on Computational Natural Language Learning" (CoNLL 2007) esplicitamente dedicata alla valutazione comparativa dei parsers a dipendenze per lingue diverse. In quell'occasione è stato infatti distribuito lo script eval07.pl[26] usato in questo lavoro per il computo dei valori di LAS, UAS e LA su AMBnorm–test e ISST–TANL–test, calcolati sia in generale (vedi Paragrafo 3.4.1) sia rispetto alla singola categoria morfosintattica del token dipendente (vedi Paragrafo 3.4.2).

Un'analisi più dettagliata dell'accuratezza di DeSR è stata inoltre condotta sulla base dei valori percentuali di 'precision' e 'recall' nell'annotazione delle singole relazioni di dipendenza. I valori sono stati calcolati in questo modo[27]:

- la 'precision' 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 trovate nel test corpus;

- 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'.

Mentre dunque il calcolo della precision permette di valutare la 'precisione' dell'analisi, cioè il numero di analisi corrette rispetto a tutte le analisi

[25]Per una descrizione dettagliata del 'task' vedi (Nivre et al., 2007a).

[26]Lo script è disponibile alla pagina http://depparse.uvt.nl/depparse-wiki/SoftwarePage#eval07.pl.

[27]Anche questi valori sono stati ottenuti usando lo script eval07.pl

realizzate, il valore di 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' corpus. Come discusso nel Paragrafo 3.4.3, il confronto di tali valori in AMB–norm–test e in ISST–TANL–test è fondamentale per individuare quali sono i tipi di relazione di dipendenza che rappresentano i maggiori ostacoli alla corretta annotazione dei testi giuridici.

3.4.1 LAS, UAS e LA generali

I risultati del confronto tra l'accuratezza di DeSR nell'annotazione dei testi normativi e di quelli giornalistici sono mostrati nella Tabella 3.2, dove sono riportati i valori di LAS, UAS e LA per le due tipologie di testi.

Come ci si aspettava il parser ha prestazioni migliori sulla tipologia di testi sui quali è stato addestrato, cioè quelli giornalistici. In particolare, la differenza tra il valore di LAS ottenuto nell'annotazione di ISST–TANL–test (80,02%) e di AMBnorm–test (74,10%) è di 5,92 punti percentuali.

Valori superiori si hanno quando la valutazione è meno restrittiva. Così, sebbene UAS e LA siano maggiori nella valutazione di ISST–TANL–test, tuttavia i valori ottenuti in AMBnorm–test sono superiori a quelli ottenuti valutando la capacità del parser di ricostruire l'intera struttura sintattica di un periodo.

	AMBnorm–test	ISST–TANL–test
LAS	74,10	80,02
UAS	76,70	84,26
LA	86,54	89,02

Tabella 3.2: LAS, UAS e UA in AMBnorm–test e in ISST–TANL–test.

3.4.2 LAS e UAS rispetto alle singole categorie morfosintattiche

Maggiori dettagli sulla valutazione delle performance di DeSR nell'annotazione dei testi normativi sono dati dai valori di LAS e UAS calcolati rispetto alle singole categorie morfosintattiche del token dipendente[28].

[28]Considerata la minore significatività del calcolo di LA, si è qui deciso di tralasciare questo dato.

La Tabella 3.3 mostra il confronto dei valori di LAS e UAS in AMBnorm–test e ISST–TANL–test, calcolati come la percentuale di volte in cui il parser ha assegnato correttamente ad un dipendente (al quale è stata precedentemente assegnata una determinata categoria morfosintattica[29]) sia la testa sintattica sia il tipo di relazione di dipendenza (LAS) o la sola testa sintattica (UAS).

	AMBnorm–test		ISST–TANL–test		
Categoria morfosintattica	LAS	UAS	LAS	UAS	Differenza
S	84	86	85	89	|1|
F	47	50	71	72	|24|
V	70	72	80	82	|10|
E	61	66	68	81	|7|
R	98	98	99	99	|1|
A	92	92	92	93	|0|
P	88	91	83	90	|5|*
B	79	81	77	83	|2|*
C	51	53	59	64	|8|
N	81	84	81	87	|0|
D	90	90	100	100	|10|
T	100	100	100	100	|0|

Tabella 3.3: LAS e UAS in AMBnorm–test e in ISST–TANL–test rispetto alla categoria morfosintattica del token dipendente.

Come si può vedere dalla differenza tra di valori di LAS, riportata nell'ultima colonna della Tabella 3.3 (colonna *Differenza*), le categorie morfosintattiche rispetto alle quali DeSR ha prestazioni peggiori in AMBnorm–test che in ISST–TANL–test sono nell'ordine i tokens di punteggiatura (F), quelli verbali (V) e i tokens determinanti (D), le congiunzioni (C) e le preposizioni (E)[30]. Rispetto a questa tipologia di tokens dipendenti lo scarto tra la valutazione della corretta identificazione della testa sintattica e del tipo di dipendenza è infatti rispettivamente di 24, 10, 8 e 7 punti percentuali.

Questi dati sono complementari a quelli visualizzati nella Figura 3.6, che mostra la diversa distribuzione percentuale in AMBnorm–test e in ISST–

[29]Per la lista completa dei tipi di categorie morfosintattiche vedi l'Allegato I.

[30]La differenza è stata calcolata in valori assoluti tra ISST–TANL–test e AMBnorm–test. Gli asterischi segnalano i casi in cui i valori di LAS di AMBnorm–test sono maggiori di quelli di ISST–TANL–test.

TANL–test degli errori commessi da DeSR valutati in termini di LAS[31]. Come si può vedere dalla distanza tra le linee che nel grafico rappresentano la diversa percentuale di errori rispetto alla categoria morfosintattica del token dipendente, la differenza maggiore tra le due tipologie di corpora valutati si ha rispetto all'erroneo riconoscimento della testa sintattica e del tipo di dipendenza di tokens di punteggiatura (con una differenza di 12 punti percentuali), di verbi (con 7 punti di scarto) e di preposizioni (con 5 punti di scarto). Inoltre, la quarta tipologia di tokens rispetto alla quale il parser commette il maggior numero di errori in AMBnorm–test è quella dei sostantivi con una percentuale di errori pari all'11%.

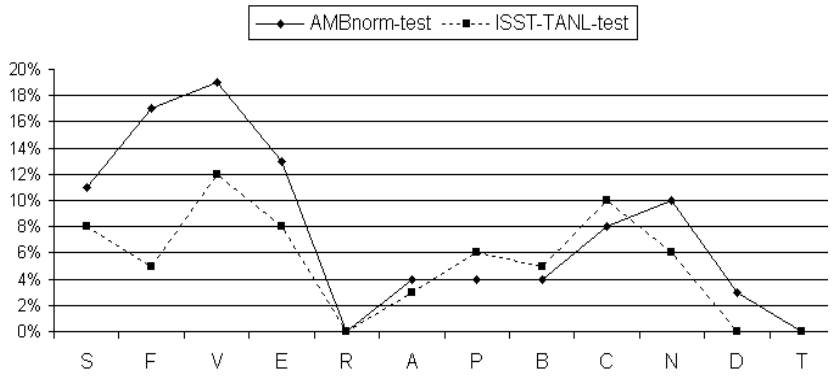


Figura 3.6: La diversa percentuale di errori ('error rate') valutati in termini di LAS in AMBnorm–test e ISST–TANL–test rispetto alla categoria morfosintattica del token dipendente.

È qui da notare che i risultati delle valutazioni sin qui condotte devono essere interpretare alla luce *i)* del grado di precision e recall raggiunto da DeSR nella ricostruzione dei singoli tipi di dipendenza (Paragrafo 3.4.3) e *ii)* delle principali caratteristiche linguistiche proprie degli atti normativi (Capitolo 4). I punteggi di LAS rispetto alle singole categorie morfosintattiche dei tokens dipendenti sono infatti, da un lato, legati all'accuratezza del parser nel corretto riconoscimento delle triple [testa, dipendente, tipo di dipendenza]; dall'altro, sono riconducibili ad alcune specificità sintattiche dei testi giuridici

[31]I valori sono quelli dell''error rate' calcolato dallo script eval07.pl. Considerato il carattere più restrittivo del calcolo del LAS, si è deciso di riportare qui l'error rate valutato solo rispetto a questo tipo di valutazione e non anche in termini di UAS e LA.

esaminati che, contenendo strutture diverse (o con diversa distribuzione d'uso) dai testi giornalistici rispetto ai quali sono confrontati, sono responsabili della diminuzione dell'accuratezza generale dell'annotazione automatica.

Ad esempio, il fatto che DeSR nella corretta assegnazione della testa sintattica e del tipo di dipendenza ai tokens (dipendenti) di tipo preposizionale abbia il 61% di LAS in AMBnorm–test, con uno scarto di 7 punti percentuali rispetto al 68% in ISST–TANL–test, è legato *i)* alla precision con cui il parser annota correttamente i tipi di ruolo sintattico–funzionale che, sulla base dei criteri di annotazione adottati, un token preposizionale può svolgere rispetto alla testa sintattica da cui dipende e *ii)* alla percentuale di occorrenza di preposizioni e sostantivi organizzati in strutture nominali complesse.

Da un alto, infatti, come mostrato nella Tabella 3.4 del Paragrafo 3.4.3, la precisione nella corretta annotazione di dipendenze di tipo 'comp, comp_ind, comp_loc, comp_temp, mod_rel, mod_loc' e 'arg' svolte da un token preposizionale diminuisce in AMBnorm–test rispetto a ISST–TANL–test. Dall'altro, su questa differenza di accuratezza influisce una ben nota peculiarità della tipologia di testi giuridici in esame, rilevata in fase di monitoraggio nel Capitolo 4. Come discusso in quell'occasione, l'occorrenza percentuale di preposizioni e sostantivi maggiore rispetto ai testi giornalistici di confronto è legata ad una caratteristica organizzazione sintattica del periodo giuridico: la spiccata propensione per la modificazione nominale articolata in lunghe catene di complementi preposizionali dipendenti da teste nominali[32]. Come dimostrano dunque i risultati della valutazione qui riportati, questo diverso comportamento sintattico è tra i principali responsabili della diminuzione di accuratezza del parser nell'annotazione di testi giuridici.

3.4.3 Precision e Recall nell'annotazione dei singoli tipi di relazione di dipendenza

I risultati di precision e recall di DeSR nell'annotazione delle singole relazioni di dipendenza[33] in ISST–TANL–test e AMBnorm–test, riportati nella Tabel-

[32]Nota che la maggiore distribuzione percentuale di preposizioni e sostantivi nei testi giuridici rispetto a quelli giornalistici è tale anche nei due test corpora usati per la valutazione di DeSR. Mentre infatti in AMBnorm–test le preposizioni sono il 22% del totale di tokens e i sostantivi sono il 28,69%, in ISST–TANL–test le prime sono il 14,42% e i secondi il 25,25%.

[33]Per la lista completa dei tipi di relazione di dipendenza vedi l'Allegato I.

la 3.4 insieme ai valori di F–Measure[34], sono interpretati e discussi in quanto segue alla luce del confronto tra il gold corpus e l'annotazione prodotta dal parser in modo automatico.

I tipi di relazione di dipendenza rispetto ai quali c'è una significativa diminuzione in termini di precision in AMBnorm–test rispetto a ISST–TANL–test costituiscono il punto di partenza delle analisi. Confrontando in particolare le annotazioni presenti nel gold e quelle prodotte automaticamente nel test corpus di atti normativi, è stato possibile così rintracciare le principali tipologie di errori di annotazione commessi.

L'obiettivo ultimo era quello di individuare strutture sintattiche tipiche dei testi normativi che non ricorrendo o ricorrendo con una frequenza minore nei testi giornalistici di riferimento sono maggiormente responsabili di una diminuzione di accuratezza nell'annotazione automatica di alcuni tipi di relazione di dipendenza.

A questo scopo è stato utilizzato MaltEval[35] (Nilsson e Nivre, 2008), uno strumento di valutazione e di visualizzazione grafica per annotazioni sintattiche a dipendenze. Esso consente di confrontare visivamente le annotazioni di due corpora. Ciò ha permesso di condurre un'analisi delle specifiche porzioni di testo nelle quali occorrono differenze di annotazione tra AMBnorm–gold e AMBnorm–test.

In quanto segue le performances di precision di DeSR nell'annotazione di AMBnorm-test sono dunque descritte facendo riferimento a comportamenti sintattici particolari responsabili della diminuzione dei valori di precision rispetto all'annotazione automatica di ISST–TANL–test.

È da notare, in primo luogo, come sia nella corretta identificazione dell'esistenza della radice sintattica ('root') dell'intero periodo che il parser mostra la differenza maggiore nell'annotazione delle due tipologie di testi. Passando dal 79,29% in ISST–TANL–test al 33,59% in AMBnorm–test, la precision diminuisce infatti di ben 45 punti percentuali. Ciò è legato alla difficoltà del parser di analizzare correttamente periodi che presentano una struttura sintattica drasticamente differente da quella dei testi giornalistici sui quali è stato addestrato, periodi dei quali DeSR non riesce ad individuare la radice.

Le principali motivazioni sottostanti alle differenze tra AMB–norm–test e ISST–TANL–test in termini di precision di annotazione delle singole relazioni

[34]Si tratta della media pesata tra precision e recall.
[35]http://w3.msi.vxu.se/∼jni/malteval/

	AMBnorm–test			ISST–TANL–test		
Tipo di relazione	Precision	Recall	F-Measure	Precision	Recall	F-Measure
root	33,59	58,11	42,57	79,29	79,86	79,57
arg	59	71,95	64,83	75,51	77,08	76,29
aux	100	96,63	98,29	98,06	96,19	97,12
clit	100	93,75	96,77	93,75	86,54	90,00
comp	67	70,33	68,62	71,73	80,24	75,75
comp_ind	20	5,56	8,70	85	60,71	70,83
comp_loc	25	15,38	19,04	53,97	39,53	45,63
comp_temp	44,44	25	32,00	61,11	29,73	40,00
con	38,89	35,9	37,34	59,8	54,59	57,08
concat	0	0	0,00	100	66,67	80,00
conj	33,04	38,19	35,43	56,47	53,33	54,86
det	98,33	98,33	98,33	99,25	99	99,12
dis	57,14	56	56,56	55,56	38,46	45,45
disj	67,65	46	54,76	50	16,67	25,00
mod	81,82	76,58	79,11	77,58	79,43	78,49
mod_loc	75	75	75,00	90,91	62,5	74,07
mod_rel	40,48	28,81	33,66	53,03	61,4	56,91
mod_temp	0	0	0,00	75,76	50	60,24
modal	93,75	100	96,77	93,94	100	96,88
neg	96,55	96,55	96,55	93,48	93,48	93,48
obj	85	91,07	87,93	81,91	89,53	85,55
pred	81,48	86,27	83,81	82,14	79,31	80,70
prep	97,35	97,19	97,27	97,16	97,69	97,42
punc	52,59	54,34	53,45	72,85	75,47	74,14
sub	75,76	65,79	70,42	85	89,47	87,18
subj	63,64	69,47	66,43	80,67	76,19	78,37
subj_pass	90,91	31,75	47,06	61,54	47,06	53,33

Tabella 3.4: Precision, Recall e F–Measure in AMBnorm–test e in ISST–TANL–test rispetto al tipo di relazione di dipendenza.

di dipendenza sono in particolare riconducibili ai seguenti comportamenti sintattici caratteristici dei testi giuridici.

***i)* La lunghezza delle relazioni di dipendenza.** Come precedemente discusso nel Paragrafo 3.3.1.3, il riconoscimento dell'esistenza di una relazione di dipendenza tra due tokens molto distanti tra loro all'interno di un periodo è un compito particolarmente complesso per i parsers a dipendenze 'data–

driven'. Tuttavia, come mostrato in fase di monitoraggio linguistico degli atti giuridici esaminati in questo studio[36], tale tipologia di testi è caratterizzata da una distanza media tra token dipendente e testa sintattica maggiore di quella presente nei testi giornalistici. Una tale peculiarità è responsabile di una diminuzione delle performances di DeSR nella corretta annotazione di diverse relazioni di dipendenza.

Un caso emblematico è quello della diminuzione della precision nel riconoscimento della relazione 'comp'. Sebbene si tratti di una differenza inferiore a quella dell'annotazione di 'root', essa passa dal 71,73% in ISST–TANL–test al 67% in AMBnorm–test. Un esempio di come ciò sia riconducibile (tra le altre cause) alla distanza tra il dipendente e la sua testa sintattica è dato dal seguente periodo:

- *Il pagamento dell'ammenda per le emissioni in eccesso non dispensa il gestore dall'obbligo di restituire un numero di quote di emissioni corrispondente a tali emissioni in eccesso all'atto della restituzione delle quote relative alle emissioni dell'anno civile seguente.*

Come risulta evidente dal confronto delle Figure 3.7 e 3.8, che riportano rispettivamente l'annotazione manuale di *restituire un numero di quote di emissioni corrispondente a tali emissioni in eccesso all'atto* in AMBnorm–gold e quella automatica in AMBnorm–test, il parser non ha riconosciuto che la testa sintattica del dipendente *all'* (30) che svolge una relazione 'comp' è *restituire* (17) e non *quote* (21). La mancata assegnazione corretta della testa è riconducibile alla distanza di 12 tokens rispetto al dipendente, una distanza considerevole tenuto conto del fatto che i testi giornalistici sono caratterizzati da una distanza media testa/dipendente di circa 8 tokens[37].

In questo caso è inoltre coinvolto il riconoscimento della testa sintattica del token *corrispondente* (24), erroneamente legato da una relazione di dipendenza 'mod' a *emissioni* (23), mentre si tratta di un modificatore del token *numero* (19), come mostrato nella Figura 3.7.

Un altro esempio è quello rappresentato dal seguente periodo:

- *I piani e i programmi il cui primo atto preparatorio formale è precedente a tale data e che sono stati approvati o sottoposti all'iter legislativo più di ventiquattro mesi dopo la stessa data sono soggetti all'obbligo di cui*

[36] Vedi in particolare il Paragrafo 4.2.3.2.
[37] Vedi in proposito il Paragrafo 4.2.3.2.

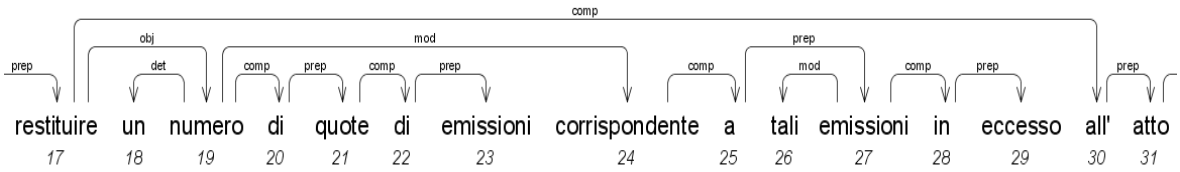


Figura 3.7: *restituire un numero di quote di emissioni corrispondente a tali emissioni in eccesso all'atto*: annotazione in AMBnorm–gold.

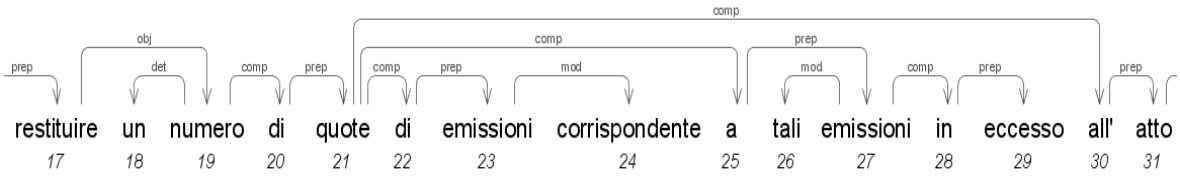


Figura 3.8: *restituire un numero di quote di emissioni corrispondente a tali emissioni in eccesso all'atto*: annotazione in AMBnorm–test.

> *all'articolo 4, paragrafo 1, a meno che gli Stati membri decidano caso per caso che ciò non è possibile, informando il pubblico di tale decisione.*

In questo caso, durante la fase di annotazione automatica non è stata correttamente assegnata al token *piani* la relazione di dipendenza 'subj' dalla testa sintattica *sono*. Ciò è principalmente dovuto alla grande distanza di 32 tokens tra testa e dipendente. È questa infatti una delle cause della diminuzione di 17 punti percentuali della precisione nell'annotazione della relazione 'subj', che passa dall'80,67% in ISST–TANL–test al 63,64% in AMBnorm–test.

Anche la diminuzione della precisione nel corretto riconoscimento delle relazioni 'arg', 'conj' e 'con' è riconducibile a questa caratteristica dei testi normativi. Non a caso, come mostra la Tabella 3.4, queste sono tre tra le relazioni di dipendenza per le quali c'è la maggiore differenza tra ISST–TANL–test e AMBnorm–test in termini di valori percentuali di precision.

Nel primo caso, un esempio è rappresentato dal seguente periodo:

- *Gli Stati membri istituiscono regimi che obblighino coloro che hanno immesso o intendono immettere sul mercato biocidi e coloro che chie-*

dono l'iscrizione di principi attivi negli allegati I, I A e I B a pagare tasse che corrispondano nella misura del possibile ai costi che essi devono sostenere ai fini dell'espletamento di tutte le diverse procedure connesse con le disposizioni della presente direttiva.

In fase di annotazione sintattica automatica il parser non ha correttamente assegnato la relazione 'arg' che lega *obblighino* (testa sintattica) e il token *a* (dipendente), a sua volta testa della frase argomentale *a pagare tasse che ....* L'errore è causato dalla grande distanza testa/dipendente pari a 28 tokens.

Nel caso del seguente periodo, la lunghezza delle relazioni 'conj' e 'con' ha fatto sì che il parser non sia riuscito ad individuarle correttamente:

- *Gli operatori di telecomunicazioni hanno l'obbligo di tenere indenne l'ente locale, ovvero l'ente proprietario, dalle spese necessarie per le opere di sistemazione delle aree pubbliche specificamente coinvolte dagli interventi di installazione e manutenzione e di ripristinare a regola d'arte le aree medesime nei tempi stabiliti dall'ente locale.*

Non è cioè stata riconosciuta l'esistenza di una relazione 'conj' tra il token *di*, testa sintattica della frase argomentale *tenere indenne l'ente locale, ovvero ...*, e il token *di*, che svolge il ruolo di dipendente ma che è a sua volta testa sintattica della frase argomentale *ripristinare a regola d'arte le aree ...* . In questo caso la distanza testa/dipendente è di 31 tokens. Per lo stesso motivo, sulla base dello schema di annotazione seguito, non è stata correttamente assegnata la relazione 'con' tra *di*[38] (testa) e il token *e* (dipendente) che coordina le due frasi argomentali.

***ii)* La presenza di frasi relative ellittiche.** In questo caso la discussione riguarda il grado di precision di DeSR nel riconoscere correttamente l'esistenza della relazione di dipendenza 'mod_rel', che passa dal 53,03% in ISST–TANL–test al 40,48% in AMBnorm–test. Si tratta di una relazione piuttosto complessa, la cui identificazione è resa ancora più difficile da una specificità dei testi normativi, quella cioè dei rimandi espliciti ad altri testi normativi o a parti dell'articolato espressi con frasi relative ellittiche.

Come discusso nel Paragrafo 3.3.1.3, data l'assenza di un tale costrutto nei testi giornalistici, in fase di costruzione del gold corpus di testi normativi

[38]Testa sintattica della frase argomentale *tenere indenne l'ente locale, ovvero ...* .

è stato infatti necessario specializzare lo schema di annotazione per trattare in modo adeguato occorrenze di questo tipo. Di conseguenza, dunque, non essendo addestrato al loro riconoscimento il parser diminuisce le sue performances di analisi nell'annotazione di frasi del tipo esemplificato dal seguente periodo:

- *Gli enti locali, che per l'esercizio di funzioni di loro competenza utilizzino le opere di bonifica di cui al presente articolo, sono chiamati a contribuire alle spese per la realizzazione, l'esercizio e la manutenzione delle stesse.*

Come si può vedere confrontando le Figure 3.9 e 3.10, che riportano rispettivamente l'annotazione manuale presente nel corpus gold e quella automatica in AMBnorm–test, DeSR non ha riconosciuto la relazione di dipendenza che lega *opere* (16) con la frase relativa *di cui al presente articolo*.

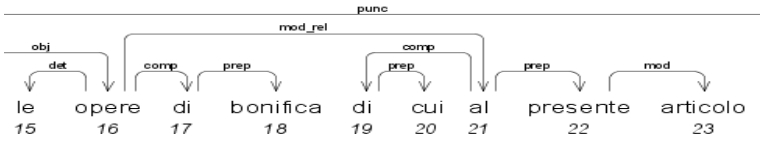


Figura 3.9: *... le opere di bonifica di cui al presente articolo*: annotazione in AMBnorm–gold.

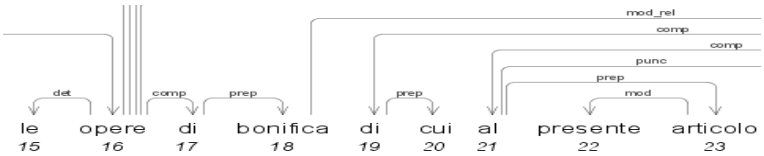


Figura 3.10: *... le opere di bonifica di cui al presente articolo*: annotazione in AMBnorm–test.

***iii)* Fenomeni di marcatezza sintattica.** Come fatto notare da Garavelli (2001, pp. 86–99), una delle peculiarità dei testi giuridici consiste nell'uso marcato dell'ordine normale soggetto–verbo–oggetto. Casi di marcatezza

sintattica di questo tipo sono riconducibili, secondo Mortara Garavelli, ad intenti pragmatici. Così ad esempio la postposizione del soggetto al verbo è dettato dalla volontà di focalizzare l'attenzione sul *nuovo* espresso dal soggetto.

Come nel caso della grande distanza testa/dipendente discussa al punti *i)*, un tale comportamento sintattico, differenziandosi da quello caratteristico di testi giornalistici, è tra le cause dell'inferiore livello di precision nella corretta identificazione automatica della relazione 'subj' in AMBnorm–test.

Un esempio è rappresentato dal seguente periodo:

- *Quando le concentrazioni superano determinate soglie di valutazione, dovrebbe essere obbligatorio un monitoraggio dell'arsenico, del cadmio, del nickel e del benzo(a)pirene.*

In questo caso, come mostrato nella Figure 3.11 e 3.12, che riportano l'annotazione di *dovrebbe essere obbligatorio un monitoraggio* nel gold corpus e quella automatica in AMBnorm–test, il parser ha riconosciuto l'esistenza di una relazione di dipendenza del token *monitoraggio* (14) dalla testa sintattica *essere* (11), ma non ne ha correttamente individuato il tipo di relazione 'subj'.

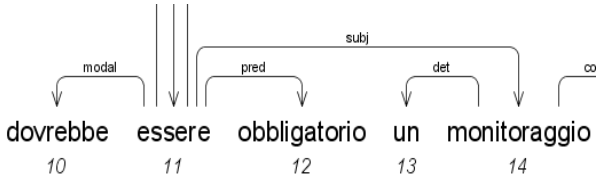


Figura 3.11: *dovrebbe essere obbligatorio un monitoraggio*: annotazione in AMBnorm–gold.

***iv)* La "densità dei vizi interpuntivi".** La questione riguarda la particolare situazione del sistema interpuntivo dei testi giuridici. Secondo quanto descritto da Garavelli (2001, p. 82) i testi normativi sono quelli nei quali si concentra la maggiore "densità dei vizi interpuntivi", che si accompagnano spesso a "difformità sintattiche".

Un uso della punteggiatura spesso non omogeneo e comunque diverso da quello dei testi giornalistici è tra le cause responsabili della diminuzione di

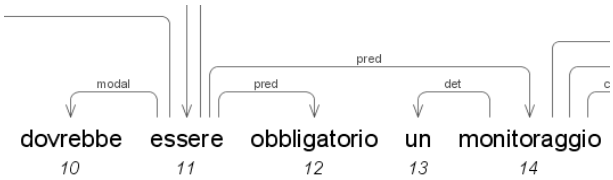


Figura 3.12: *dovrebbe essere obbligatorio un monitoraggio*: annotazione in AMBnorm–test.

20 punti percentuali nella corretta identificazione automatica della relazione 'punc' in AMBnorm–test, che passa dal 72,85% (in termini di precision) in ISST–TANL–test al 52,59%.

3.5 Verso l'adattamento di strumenti di trattamento automatico del linguaggio per l'annotazione sintattica di testi giuridici

La centralità dei paragrafi precedenti è legata alla constatazione che quanto descritto rappresenta il punto di partenza per uno studio futuro volto a definire una metodologia di adattamento di un parser a dipendenze, basato su di un approccio 'data–driven', all'annotazione di corpora di testi normativi. In questo senso, le potenzialità del lavoro sin qui svolto sono di due tipi.

A differenza degli studi 'grammar–driven' finalizzati ad adattare strumenti di annotazione sintattica automatica all'elaborazione di testi di dominio estendendo una grammatica sviluppata per la lingua comune, negli studi basati su di un approccio 'data–driven' il quesito al quale si cerca di trovare una risposta è il seguente: "Can training data from one corpus be applied to parsing another?".

È infatti questa la domanda posta da Gildea (2001), lo studio considerato pioniere nelle ricerche volte alla definizione di una metodologia per adattare parsers statistici all'annotazione di testi caratterizzati da un dominio diverso da quello del 'training' corpus sul quale sono stati addestrati. Partendo dall'osservazione che al cambiare del dominio di conoscenza del testo analizzato l'accuratezza del parser diminuisce notevolmente, l'obiettivo condiviso è

quello di sviluppare un algoritmo di elaborazione del testo in grado di essere ugualmente accurato nell'annotazione di testi di nuovi domini.

Come riportato da Gildea (2001), infatti, anche il solo cambiare la tipologia di linguaggio comune contenuto in corpora di testi giornalistici analizzati inficia le performances del parser. Così il Collins'parser (Collins, 1999) addestrato su testi del Wall Street Journal diminuisce la propria accuratezza di analisi di 5,5 punti percentuali quando testato sul Brown Corpus, passando dall'86,6% all'81% di precision. Lo stesso parser, inoltre, come dimostrato da Clegg e Shepherd (2005), addestrato sulla Penn TreeBank (Marcus et al., 1993) e testato sul GENIA corpus[39], un corpus di abstracts di articoli biomedici, diminuisce di 7,7 punti percentuali, passando dall'86,8% al 79% di precision.

Il problema è ancor più di rilievo se si tiene in considerazione il fatto che la fase di annotazione sintattica di un testo costituisce il punto di partenza per numerose applicazioni pratiche, quali ad esempio l'estrazione automatica di informazione, la traduzione automatica, Question Answering, ecc... Non a caso infatti, il 'sublanguage' oggetto della maggior parte dei lavori in materia è il linguaggio biomedico, linguaggio caratterizzante il dominio di conoscenza per il quale c'è un grande interesse per lo sviluppo di applicazioni di 'Text Mining'[40]. Anche se tuttavia le più recenti attività sono rivolte alla definizione di metodologie in grado di adattare gli strumenti di analisi all'elaborazione di collezioni testuali eterogenee rispetto al dominio (McClosky et al., 2010). Questo è in linea con la necessità di annotare le grandi e variegate quantità di testi presenti nel web.

Inoltre, sempre più oggi l'approccio 'data–driven' all'elaborazione del testo rappresenta lo stato dell'arte degli strumenti usati in un compito di adattamento al dominio. Come hanno recentemente fatto notare infatti Plank e van Noord (2010), "only few studies examined the adaptation of *grammar–based* systems". Questo è testimoniato dal fatto che i più recenti studi in materia presentati in occasione del "Workshop on Domain Adaptation for Natural Language Processing" (DANLP 2010)[41] sono tutti basati su strumenti statistici, basati su algoritmi di apprendimento automatico.

[39] http://www-tsujii.is.s.u-tokyo.ac.jp/∼genia/topics/Corpus/

[40] Alcuni dei più recenti e rilevanti lavori sono stati svolti da Lease e Charniak (2005); Nivre et al. (2007b); McClosky e Charniak (2008); Plank e van Noord (2011).

[41] Per una rassegna dei contributi vedi gli atti dell'edizione 2010 del disponibili alla pagina http://aclweb.org/anthology-new/W/W10/W10-2600.pdf

Tuttavia, come discusso nel Paragrafo 3.1, sino ad oggi il limite maggiore di questo tipo di strumenti è quello di essere legati alla presenza di un 'training' corpus di dominio sul quale addestrare gli strumenti di analisi al riconoscimento di testi caratterizzati da un linguaggio diverso da quello del 'training' corpus originario. Si tratta di quello che viene comunemente definito 'supervised domain adaptation scenario', il caso in cui cioè sia richiesta una quantità, anche ridotta, di testi annotati in modo manuale per adattare gli strumenti di analisi ad un nuovo dominio.

In questo senso, dunque, il lavoro di costruzione di un 'gold' corpus di atti normativi annotati fino al livello sintattico è il punto di partenza per la raccolta di una più ampia collezione di testi di questo tipo da usare in fase di addestramento di un parser a dipendenze. È d'interesse qui ricordare che un corpus di questo tipo rappresenta una delle rare eccezioni nell'ambito delle attività di ricerca in materia di gestione dell'informazione giuridica basate su strumenti e metodi di Trattamento Automatico del Linguaggio. Come ricordato infatti nel Paragrafo 2.3.2.1, sebbene se ne senta la necessità[42], la costruzione di 'gold' corpora di testi giuridici da usare in fase di addestramento automatico degli strumenti statistici di annotazione linguistica è sino ad oggi oggetto di poche attenzioni. Due eccezioni significative sono quelle rappresentate da Walter (2009) per la lingua tedesca e dalla porzione della TUT costituita da articoli del Codice Civile italiano.

Data la sua unicità, AMBnorm–gold è tra i dati messi a disposizione dei partecipanti al "Domain Adaptation for Dependency Parsing task"[43] dell'edizione 2011 di Evalita attualmente in corso. In questo contesto il corpus è utilizzato per lo svolgimento di una delle due parti in cui è articolato il 'task', quello che consiste cioè nello sviluppare un algoritmo di annotazione sintattica a dipendenze facendo affidamento su di un 'training' corpus formato da testi di lingua comune e da una ridotta quantità di testi di dominio.

Il secondo vantaggio dello studio condotto nei paragrafi precedenti è quello legato ai risultati e alle discussioni sul diverso grado di accuratezza di DeSR nell'annotazione di corpora di testi giornalistici e di atti normativi. Come fatto osservare da Clegg e Shepherd (2005), una delle attività preliminari alla definizione di una metodologia di adattamento degli strumenti di trattamento automatico del testo è quella relativa all'analisi degli errori. Una tale analisi può infatti essere d'aiuto "to identify both the sources of performances

[42] Vedi McCarty (2009).

[43] http://www.evalita.it/2011/tasks/dependency_parsing

problems and to a certain extent their causes and connotations".

In questo senso, pertanto, non solo le osservazioni relative alla generale diminuzione (in termini di LAS e UAS) dell'accuratezza del parser su AMBnorm–test rispetto alla tipologia di testi del 'training' corpus contribuiscono a definire quantitativamente l'impatto che la lingua del diritto qui in esame ha sulle performances di analisi. Sono centrali per lo sviluppo di una futura strategia di adattamento degli strumenti anche le analisi specifiche dei tipi di relazione di dipendenza maggiormente coinvolti nella riduzione della precisione di annotazione.

È infine qui importante mettere in luce come anche la definizione di una metodologia di monitoraggio delle caratteristiche linguistiche di testi giuridici descritta nel Capitolo 4 sia finalizzata ad offrire un supporto ad attività di adattamento degli strumenti di annotazione del testo al dominio giuridico. Come fatto osservare infatti da Gildea (2001), lo studio delle variazioni linguistiche in corpora rappresentativi varietà linguistiche diverse è di fondamentale importanza nella progettazione e sviluppo di parsers statistici. Sottolineando, in particolare, l'importanza degli studi sulle variazioni tra registri realizzati da Douglas Biber, Gildea chiarisce che "the frequencies of various structures in training data are reflected in a statistical parser's probability model". È importante qui ricordare come Biber (1993) stesso, indirizzando esplicitamente i risultati delle sue ricerche a questa comunità di ricerca, affermi che le differenze da lui raccolte tra registri linguistici "are also important for probabilistic part–of–speech taggers and syntactic parsers, because the probabilities associated with grammatically ambiguous forms are often markedly different across registers".

3.6 Considerazioni conclusive

Il contributo più originale di questo capitolo riguarda l'attenzione posta sulla valutazione quantitativa dell'impatto che la lingua del diritto ha sull'accuratezza delle analisi prodotte da strumenti di Trattamento Automatico del Linguaggio di tipo 'data–driven' addestrati su testi giornalistici rappresentativi della lingua comune.

Come messo in luce nel Paragrafo 2.3.2.2, pochi dei più recenti studi finalizzati all'utilizzo di strumenti di annotazione linguistica del testo per la gestione automatica del contenuto di corpora di documenti giuridici di diverso genere affrontano il tema in maniera esaustiva. Inoltre, nessuno di

essi tratta il caso della lingua italiana. In questo senso, dunque, i risultati dei vari tipi di valutazione condotti nei paragrafi precedenti rappresentano una novità sia per la comunità di ricerca in materia di AI&Law sia per quella di linguistica computazionale.

L'impatto e i vantaggi dello studio condotto in questo capitolo possono essere infatti ricondotti a tre tipologie. Da un lato, la costruzione di un 'gold' corpus di atti normativi linguisticamente annotato in modo manuale fino al livello sintattico è d'interesse per le ricerche in materia di adattamento di strumenti di Trattamento Automatico del Linguaggio ad un dominio specialistico. Come ricordato, gli ostacoli posti dall'elaborazione di testi caratterizzati da un linguaggio diverso da quello per il quale gli strumenti sono stati costruiti sono da sempre al centro degli studi di chi progetta e sviluppa algoritmi di Trattamento Automatico del Linguaggio. In particolare, nel caso degli algoritmi di annotazione sintattica di tipo 'data–driven', avere a disposizione un 'gold' corpus di testi di un dominio diverso da quello sul quale essi sono stati addestrati è utile *i)* per valutare le performances degli algoritmi al cambiare della tipologia di linguaggio da trattare e *ii)* per essere usata come risorsa di riferimento per successivi studi di adattamento degli strumenti di annotazione ad un dominio specifico (caratterizzato da un determinato 'sublanguage').

La costruzione di AMBnorm–gold va infatti in questa direzione. Essa ha permesso *i)* di individuare le specializzazioni dei criteri di annotazione seguiti per la lingua comune, specializzazioni necessarie per generare analisi corrette di strutture sintattiche caratteristiche dei testi normativi e *ii)* di quantificare l'accuratezza dell'annotazione sintattica a dipendenze realizzata da strumenti 'data–driven' attraverso il confronto con i risultati delle analisi dei testi giornalistici sui quali gli strumenti sono stati addestrati.

È interessante qui far notare che i casi di annotazione in cui gli strumenti hanno dimostrato di avere una maggiore diminuzione di performance sono *i)* quelli che riguardano strutture sintattiche specifiche dei testi normativi per le quali in fase di costruzione di AMBnorm–gold è stata prevista una specializzazione dei criteri di annotazione, come ad esempio il caso delle frasi relative ellittiche, e *ii)* quelli che riguardano caratteristiche linguistiche che in fase di monitoraggio comparativo tra testi giornalistici e atti normativo–amministrativi[44] sono risultate essere peculiarità di questi ultimi. È il caso quest'ultimo, ad esempio, della lunghezza delle relazioni di dipendenza che,

[44]Vedi Capitolo 4.

di gran lunga maggiore negli atti normativo–amministrativi, si è rivelata una delle principali cause dell'erroneo riconoscimento dell'esistenza di una relazione tra un token dipendente e la sua testa sintattica.

Dall'altro, uno studio quantitativo focalizzato in particolate sull'impatto della lingua del diritto sull'accuratezza dell'annotazione sintattica a dipendenze, consentito dalla costruzione di AMBnorm–gold, è di fondamentale importanza per compiti di gestione automatica del contenuto di testi giuridici. Come ricordato da Nivre (2006, p. 5), a differenza di altri tipi di applicazioni "if we move to applications that require some kind of semantic analysis of individual sentences, the role of parsing becomes more evident". Ed è in discussione soprattutto la capacità di soddisfare i quattro requisiti fondamentali[45] che un parser deve possedere per poter generare analisi affidabili come punto di partenza per il successivo livello di annotazione semantica.

In questo senso, dunque, lo studio condotto nei precedenti paragrafi è preliminare a quello esposto nella Parte III di questo lavoro, dove la fase di annotazione sintattica a dipendenze è stata considerata il punto di partenza per l'annotazione semantica di testi normativi. Sino ad oggi tale annotazione incrementale presuppone la revisione manuale delle analisi sintattiche generate in modo automatico. Ciononostante, la definizione di un processo semi o completamente automatico di annotazione semantica basata sull'output dell'annotazione sintattica (automatica) è tra le applicazioni future di una metodologia di adattamento di un parser all'analisi di testi giuridici.

[45]I quattro requisiti sono: "robustness", "disambiguation", "accuracy" e "efficiency".

Capitolo 4

Il monitoraggio delle caratteristiche linguistiche di testi giuridici

Questo capitolo ha l'obiettivo di dimostrare come i risultati dell'annotazione linguistica automatica, pur contenendo inevitabilmente un margine di errore, ulteriormente accentuato dalle specificità della lingua del diritto e dalle sue difficoltà di analisi, se appropriatamente esplorati possono fornire indicazioni affidabili per la descrizione delle principali caratteristiche linguistiche di un testo giuridico.

Allo scopo pertanto di tracciare il profilo linguistico degli atti normativi e amministrativi contenuti nel corpus di testi giuridici qui preso in esame, è stata messa a punto una metodologia di analisi finalizzata a descriverne le caratteristiche lessicali, morfosintattiche e sintattiche sulla base del modo in cui alcuni significativi tratti linguistici si distribuiscono nei testi. Come dimostrano i risultati ottenuti, ciò ha permesso di fornire una serie di dimostrazioni empiriche di quanto fatto osservare negli studi linguistici tradizionalmente condotti con metodi manuali di indagine.

Elemento chiave dell'intero capitolo è il punto di vista sia esterno sia interno da cui si è scelto di guardare alla lingua del diritto. Da un lato, infatti, le analisi si sono concentrate sul confronto tra le caratteristiche della lingua del diritto e quelle della lingua comune. L'obiettivo era quello di suggerire una possibile risposta 'operativa' alla dibattuta e aperta questione circa i non lievi problemi di delimitazione tra le due. In questo senso, le analisi sono state guidate dall'intento di dimostrare in che modo e fino a che punto

caratteristica della lingua del diritto sia quella di essere contraddistinta da "usi diversi, rispetto alla norma dell'italiano comune" rintracciabili "all'interno della selezione di elementi grammaticali dell'italiano comune", come affermato da Rovere (2005, p. 242).

Sulla scia dello studio di Giovanni Rovere, anche in questo lavoro sono stati scelti come rappresentativi della lingua comune testi giornalistici. A differenza di Rovere, tuttavia, sono stati qui presi come corpora di riferimento due collezioni di testi caratterizzati da due diverse varietà di lingua giornalistica: quella di un quotidiano ad ampia tiratura come "La Repubblica" e quella di "Due Parole", un giornale scritto con una lingua giornalistica volutamente semplificata per essere compresa da persone con un basso livello di scolarizzazione o con disabilità cognitive. Questo ha permesso di verificare fino a che punto la lingua dei testi giuridici si differenzi, da un lato, da quella usata in testi comuni che dovrebbero essere leggibili ad un ampio pubblico di lettori e, dall'altro, da quella pensata per essere estremamente semplice e comprensibile.

Dall'altro, la metodologia comparativa di analisi adottata ha permesso di focalizzare l'attenzione su come i vari tratti linguistici si distribuiscono diversamente nei diversi tipi di testi giuridici presi in esame. Ciò ha permesso di portare l'attenzione sul carattere "multiforme e complesso" (Cortelazzo, 1997) della lingua del diritto, evidenziando affinità e differenze, ad esempio, tra decreti ministeriali e ordinanze, tra atti statali e comunitari, tra leggi e la Costituzione italiana, ecc...

In definitiva, l'ottica comparativa assunta fa sì che l'intero processo di analisi condotto si configuri come un processo di **monitoraggio linguistico** di varietà linguistiche, da un lato, e di varietà testuali diverse, rappresentative di sottovarietà della lingua del diritto, dall'altro. In questo senso, l'approccio è stato ispirato dalla prospettiva di indagine di Douglas Biber e del suo gruppo di ricerca, finalizzata allo studio delle specificità linguistiche proprie di una data varietà (o registro) della lingua standard a partire dall'analisi della diversa distribuzione d'uso di tratti lessicali e grammaticali rilevanti in più corpora testuali.

Come chiarito in quanto segue, i presupposti su cui si basa la metodologia di monitoraggio linguistico messa a punto in questo studio la rendono affidabile *i)* per condurre indagini quantitative del profilo linguistico di testi giuridici e *ii)* in uno scenario applicativo, come punto di partenza per lo sviluppo di uno strumento di monitoraggio della redazione di atti normativi e amministrativi "chiari, semplici e comprensibili" e di un indicatore del loro

livello di leggibilità.

I successivi paragrafi sono dunque organizzati in questo modo: è prima di tutto presentata in dettaglio nel Paragrafo 4.1 la metodologia di monitoraggio linguistico, attraverso la descrizione dei tratti linguistici monitorati e dei corpora di testi giuridici e giornalistici presi in esame. La discussione dei risultati ottenuti, condotta nel Paragrafo 4.2, mira a dimostrare come il monitoraggio sia affidabile per individuare le caratteristiche linguistiche dell'intero corpus testi giuridici rispetto a quello di testi giornalistici, da un lato, e delle varie tipologie di atti normativi e amministrativi, dall'altro. Il Paragrafo 4.3 è infine dedicato a tracciare alcune considerazioni conclusive, mettendo l'accento su *i)* come la metodologia di monitoraggio linguistico si sia rivelata affidabile per ricostruire il profilo linguistico dei testi giuridici esaminati (Paragrafo 4.3.1) e su *ii)* come essa ponga le basi per il futuro sviluppo di uno strumento a supporto delle attività di controllo e di verifica della buona redazione di atti normativo–amministrativi e di uno strumento in grado di definirne il livello di leggibilità sulla base della distribuzione di tratti linguistici (Paragrafo 4.3.2).

4.1 La metodologia di monitoraggio linguistico

L'approccio al monitoraggio linguistico di testi giuridici seguito in questo studio parte da due considerazioni preliminari. In primo luogo, l'idea che a partire dall'annotazione linguistica automatica di un testo possano essere ricavate "indicazioni utili in merito alla definizione di strumenti di rilevazione di tipo quantitativo finalizzati alla ricostruzione del profilo linguistico di un testo", come empiricamente dimostrato da Dell'Orletta e Montemagni (2010a). Le caratteristiche linguistiche individuate nei diversi corpora sono infatti il risultato dell'analisi della diversa distribuzione d'uso di alcuni dei tratti linguistici rintracciati nei testi sulla base dell'elaborazione linguistica automatica condotta con gli strumenti di Trattamento Automatico del Linguaggio descritti nel Capitolo 3.

In secondo luogo, la metodologia di monitoraggio linguistico prende le mosse da alcune intuizioni alla base dei lavori di Douglas Biber: *i)* il fatto che "a complete description of the language often entails a composite analysis of features" (Biber, 1993, p. 220), che *ii)* "linguistic features from all

levels function together as underlying dimensions of variation, and that there are systematic and important linguistic differences among registers with respect to these dimensions" (Biber, 1993, p. 220–221) e che, dunque, *iii)* un approccio comparativo all'analisi di diverse varietà testuali sia in definitiva finalizzato a trovare una risposta alla domanda "what does 'common' or 'rare' signify?" (Biber et al., 1998, p. 8).

Sulla scia di queste osservazioni, l'analisi comparativa dei tratti linguistici rintracciati nei corpora qui in esame spazia su più livelli di descrizione linguistica. Ciò consente, da un lato, di restituire un articolato profilo linguistico dei testi e, dall'altro, di monitorare come le similarità e differenze tra i loro profili corrispondano a uno o più tratti specifici. Come dimostrano infatti i risultati del monitoraggio, tipologie di testi che si differenziano, ad esempio, per caratteristiche relative a tratti sintattici sono invece accomunati da caratteristiche lessicali. L'intento è quello di dimostrare empiricamente come la domanda posta da Fiorelli (2008) a proposito del posto occupato dalla lingua del diritto in una classificazione di linguaggi specialistici rispetto alla lingua comune non possa che avere una risposta quanto mai articolata.

Inoltre, uno dei tratti caratteristici della metodologia qui descritta consiste nel confrontare i diversi tipi di testi giuridici con due tipologie di testi giornalistici rappresentativi di due diverse varietà della lingua comune, una ampiamente comprensibile ai più e una pensata per avere caratteristiche di semplicità di lettura. Come anticipato nell'introduzione a questo capitolo e discusso nei paragrafi che seguono, ciò ha importanti ricadute sia teoriche sia applicative.

In quanto segue, sono pertanto descritti i passi fondamentali che hanno portato al monitoraggio: *i)* la scelta dei tratti linguistici da monitorare e *ii)* la raccolta dei corpora di testi giuridici e giornalistici analizzati.

4.1.1 I tratti linguistici monitorati

La scelta dei tratti linguistici considerati in fase di monitoraggio è stata condotta tenendo in considerazione due aspetti. Il primo aspetto è chiaramente esposto da Dell'Orletta e Montemagni (2010a) ed è legato alla generale affidabilità di un metodo di monitoraggio linguistico condotto a partire dai risultati di un'annotazione linguistica automatica del testo.

Esso riguarda infatti la loro "computabilità su larga scala e in modo affidabile mediante tecnologie linguistico–computazionali". Tenuto in considerazione l'impatto della lingua del diritto sugli strumenti di Trattamento

Automatico del Linguaggio, un tale aspetto è particolarmente centrale in questo studio e non va sottovalutato. Per questo motivo è importante qui ricordare che i risultati del monitoraggio discussi dei paragrafi successivi devono essere letti in relazione al loro grado di accuratezza nell'annotazione dei testi giuridici.

Il secondo aspetto tenuto in considerazione riguarda una delle potenzialità dei risultati delle analisi qui condotte, il fatto cioè di poter essere utilizzati per verificare in che misura gli atti normativi e amministrativi monitorati siano stati scritti in un linguaggio "chiaro, semplice e comprensibile". È questo infatti il suggerimento generale contenuto nella "Guida per la redazione degli atti amministrativi. Regole e suggerimenti" 2011[1].

Oggi la "Guida" raccogliendo e aggiornando la "Direttiva sulla semplificazione del linguaggio dei testi amministrativi" del Ministero della Funzione Pubblica, emanata nel maggio del 2002[2], e il "Manuale di Regole e suggerimenti per la redazione dei testi normativi", adottato dalle Regioni italiane[3], costituisce la raccolta più completa delle caratteristiche linguistiche, relative a morfologia, sintassi e lessico, che un atto normativo–amministrativo rispondente a criteri di "chiarezza, precisione, uniformità, semplicità, economia" deve avere. Si è scelto pertanto di condurre il monitoraggio dei testi giuridici tenendo in considerazione anche alcuni di quei tratti linguistici che un testo normativo–amministrativo redatto secondo le regole e i suggerimenti forniti dovrebbe contenere.

Inoltre, i tratti monitorati sono tra quelli già sperimentati con successo da Montemagni (2010) per il monitoraggio della lingua italiana nelle sue varietà diamesiche, diafasiche e diastratiche, da Dell'Orletta e Montemagni (2010a) per la valutazione delle competenze linguistiche di studenti in ambito scolastico e da Dell'Orletta et al. (2010b) per il monitoraggio del profilo linguistico di apprendenti l'italiano come L2 attraverso l'analisi delle loro produzioni scritte e dei materiali didattici loro offerti nella scuola primaria e secondaria.

A seconda del livello di informazione linguistica fornita, i tratti sono stati classificati nelle seguenti tipologie:

[1]La "Guida" è navigabile e scaricabile alla pagina
http://www.pacto.it/content/view/416/48/

[2]http://www.maldura.unipd.it/buro/dir8mag2002.html

[3]L'edizione 2007 del Manuale è consultabile alla pagina
http://www.consiglioregionale.piemonte.it/labgiuridico/dwd/manuale_oli_2008.pdf

- **tratti generali**: rintracciati sulla base del livello di segmentazione del testo in frasi e di tokenizzazione, permettono di mettere in luce caratteristiche generali del testo, quali la lunghezza media dei periodi e delle parole contenute in un corpus;

- **tratti morfosintattici**: rintracciati sulla base del livello di annotazione morfosintattica, permettono di mettere in luce le caratteristiche morfosintattiche del corpus, quali la distribuzione delle varie categorie morfosintattiche. In questo caso ci si è in particolare concentrati su

  - il rapporto tra la distribuzione di sostantivi e verbi,
  - la distribuzione di preposizioni,
  - il rapporto tra la distribuzione di congiunzioni coordinanti e subordinanti;

- **tratti sintattici**: rintracciati sulla base del livello di annotazione sintattica a dipendenze, permettono di mettere in luce le caratteristiche relative alla struttura sintattica di ogni periodo nel corpus, quali

  - la distribuzione dei vari tipi di relazioni di dipendenza, la loro lunghezza e il loro livello di incassamento gerarchico nell'albero sintattico di un periodo,
  - le dipendenze di predicati verbali,
  - le forme della modificazione nominale,
  - le forme della subordinazione e, in particolare, la distribuzione media delle frasi per periodo, la proporzione di principali e subordinate e la proporzione di subordinate implicite e esplicite;

- **tratti lessicali**: rintracciati sulla base del livello di lemmatizzazione e annotazione morfosintattica, permettono di mettere in luce le caratteristiche lessicali del corpus, quali

  - il livello di 'densità lessicale', il rapporto cioè tra lessico referenziale e funzionale,
  - il livello di varietà lessicale, attraverso il calcolo del rapporto tipo/unità (Type/Token Ratio),

– il calcolo della percentuale di parole appartenenti al Vocabolario di Base del "Grande dizionario italiano dell'uso" (De Mauro, 2000) e della loro distribuzione rispetto ai repertori di uso (Fondamentale, Alto uso, Alta disponibilità).

4.1.2 I testi giuridici monitorati

Il corpus di testi giuridici oggetto di indagine linguistica in questo studio è stato costruito nell'ambito della tesi di laurea specialistica di chi scrive[4], dove è stato linguisticamente studiato a livello lessicale, morfosintattico e sintattico 'superficiale'[5].

Si tratta di un insieme composito di atti normativi e amministrativi in materia ambientale emessi da tre diverse autorità nel periodo dal I semestre 1997 al II semestre 2005, classificati sulla base della ripartizione suggerita da Garavelli (2001, pp. 26–34)[6]. L'intero corpus è stato reperito dalla banca dati del Bollettino Giuridico Ambientale (BGA)[7], edito dall'Assessorato all'ambiente della Regione Piemonte e reso disponibile on–line dal Sistema di Documentazione Ambientale (SDA)[8], realizzato nell'ambito delle attività del Sistema Informativo Regionale Ambientale dalla Regione Piemonte che consente di accedere sia ai documenti normativi e amministrativi, sia alla produzione bibliografica su tematiche ambientali.

[4]Vedi G. Venturi, "L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale", (manoscritto) dicembre 2006.

[5]Per i risultati delle analisi linguistiche condotte vedi Venturi (2010) e Lenci et al. (2009).

[6]In particolare, AMBnorm(Stato) è composto da una collezione di decreti ministeriali, leggi, decreti legislativi, decreti del Presidente del Consiglio dei Ministri, decreti del Presidente della Repubblica, decreti legge e decreti interministeriali; AMBnorm(Regione) da una collezione di leggi regionali; AMBnorm(Europa) da una collezione di decisioni, direttive e regolamenti; AMBamm(Stato) da una collezione di deliberazioni, circolari ministeriali, accordi, ordinanze, comunicati ministeriali, direttive del Presidente del Consiglio dei Ministri e direttive ministeriali; AMBamm(Regione) da una collezione di deliberazioni della giunta regionale, determinazioni dirigenziali, circolari del presidente della giunta regionale, decreti del presidente della giunta regionale, deliberazioni del consiglio regionale e comunicati; AMBamm(Europa) da una collezione di raccomandazioni e comunicazioni.

[7]http://extranet.regione.piemonte.it/ambiente/bga/index.htm

[8]http://sda.regione.piemonte.it/

Oltre a questo insieme di testi, del corpus fa anche parte la Costituzione italiana, che si è deciso qui di analizzare nella sua versione originaria del 1947 sulla scia dello studio linguistico condotto da De Mauro (2006).

Nome del corpus	**Funzione del testo**	**Autorità emittente**	**No. di tokens**
AMBnorm(Stato)	Normativa	Stato italiano	744.064
AMBnorm(Regione)	Normativa	Regione Piemonte	112.474
AMBnorm(Europa)	Normativa	Comunità europea	453.328
AMBamm(Stato)	Amministrativa	Stato italiano	107.240
AMBamm(Regione)	Amministrativa	Regione Piemonte	182.213
AMBamm(Europa)	Amministrativa	Comunità europea	17.951
COST	Normativa	Stato italiano	10.487
Totale			**1.627.757**

Tabella 4.1: Tabella riassuntiva dei corpora di testi normativo–amministrativi analizzati.

Come mostrato nella Tabella 4.1, che ne riporta la suddivisione interna per 'funzione' e 'autorità emittente', si tratta di un corpus di 1.627.757 tokens.

4.1.3 I corpora di lingua comune usati per il confronto

I testi giuridici presi in esame sono stati confrontati in fase di monitoraggio con due corpora rappresentativi della lingua comune. A questo scopo sono state selezionate due collezioni di testi, entrambe di prosa giornalistica ma con caratteristiche diverse (vedi Tabella 4.2).

Si tratta del corpus composto da articoli di giornale estratti dal quotidiano "La Repubblica" (d'ora in avanti chiamato 'Rep'), porzione del corpus CLIC–ILC (Marinelli et al., 2003), e del corpus "Due Parole" (d'ora in avanti chiamato '2Par'), periodico di "facile lettura" costituito da testi scritti in una "lingua molto chiara, semplice e precisa" esplicitamente rivolti "alle persone che hanno bisogno di testi informativi molto leggibili e comprensibili", come si può leggere sul sito dove il corpus è liberamente consultabile.

Come precedentemente anticipato, la prosa giornalistica è stata considerata, sulla scia dello studio di Rovere (2005), come significativo termine di paragone rispetto al quale confrontare le caratteristiche linguistiche rintracciate nei testi giuridici. A differenza di Rovere, tuttavia, si è deciso di indagare le differenze della lingua del diritto non solo rispetto alla lingua contenuta in testi che dovrebbero essere leggibili da un ampio pubblico di

Nome del corpus	Fonte	No. di tokens
Rep	"La Repubblica 2002", sezione del Corpus di Lingua Italiana Contemporanea (CLIC–ILC, (Marinelli et al., 2003))	2.742.478
2Par	"Due Parole. Mensile di facile lettura", disponibile alla pagina http://www.dueparole.it/	72.987

Tabella 4.2: Tabella riassuntiva dei corpora di testi giornalistici analizzati.

lettori, come gli articoli del quotidiano "La Repubblica", ma anche rispetto a testi volutamente scritti per essere di semplice lettura e comprensione.

È questo il motivo per cui è stato scelto di confrontare il corpus di testi giuridici anche con la prosa giornalistica semplificata del corpus 2Par. Il mensile nasce infatti dagli studi di linguisti e pedagoghi condotti a partire dalla metà degli anni '80, facenti capo al GULP (il Gruppo Universitario Linguistico Pedagogico) dell'Università La Sapienza di Roma e indirizzati a "scrivere testi ex novo e secondo regole esplicite, via via definite e tarate sulle caratteristiche del destinatario" (Piemontese, 1996, p. 218). Con l'intento di fornire testi in grado di essere letti "da parte di persone con ritardo mentale" o con un basso livello di alfabetizzazione, l'impegno nella redazione di 2Par è stato dunque quello di scrivere (e in alcuni casi riscrivere) testi giornalistici e di tipo informativo sulla base di criteri di leggibilità e comprensibilità. Sono questi i criteri stabiliti da Lucisano e Piemontese (1988) nell'ambito della definizione dell'indice Gulpease, la formula per la predizione della difficoltà/semplicità di testi in lingua italiana sviluppata dal GULP.

L'obiettivo era quello di verificare quali caratteristiche linguistiche condividessero con 2Par testi di legge o della Pubblica Amministrazione, testi che, in principio, dovrebbero essere leggibili ad un pubblico di cittadini variegato rispetto al livello di istruzione e alle capacità cognitive di comprensione.

Infine, è qui d'interesse ricordare che, come dimostrano i risultati del monitoraggio delle caratteristiche linguistiche di questi due corpora condotti da Dell'Orletta e Montemagni (2010a) e da Montemagni (2010), Rep e 2Par sono due corpora con profili linguistici simili, ma sotto certi aspetti anche diversi. In entrambi i casi si tratta di corpora di prosa giornalistica, dunque di testi di tipo informativo. Tuttavia essi occupano posizioni diverse in un immaginario continuum di semplicità/complessità testuale. Mentre infatti 2Par, per i principi di semplicità e leggibilità sui quali è stato costruito, si

pone all'estremo di questo continuum, rappresentando il polo di 'semplicità', Rep occupa una posizione opposta, proponendosi come collezione di testi di 'più difficile lettura'.

Come verrà discusso nei Paragrafi 4.3.2 e 4.3.1 a conclusione di questo capitolo, ciò ha importanti conseguenze sia in fase di analisi delle similarità/differenze *i)* tra la lingua del diritto e quella comune e *ii)* tra le diverse varietà di lingua del diritto sia nella definizione di un indice di leggibilità testuale basato sul monitoraggio linguistico.

4.2 I risultati del monitoraggio

Il monitoraggio dei corpora analizzati rispetto ai tratti linguistici selezionati ha permesso di metterne in luce le più significative caratteristiche linguistiche. In quanto segue, tali caratteristiche sono esposte e discusse tenendo in considerazione tre aspetti:

- il livello di annotazione linguistica da cui derivano i tratti monitorati,
- il confronto tra la lingua del diritto e la lingua comune,
- il confronto tra le varie tipologie di testi giuridici considerati.

Inoltre, ogni volta che è stato possibile, i risultati del monitoraggio sono stati confrontati con quelli ottenuti negli studi linguistici condotti in modo manuale, mettendo in luce i casi in cui essi coincidono. Come si potrà apprezzare in quanto segue, ciò ha permesso di focalizzare l'attenzione sul fatto che in molti casi le analisi quantitative realizzate in questo studio costituiscono una conferma delle analisi precedentemente condotte. Questo a riprova di come l'uso di strumenti di Trattamento Automatico del Linguaggio per l'annotazione linguistica automatica di testi giuridici, sebbene vada inevitabilmente incontro ad errori, sia un punto di partenza affidabile per ricostruirne un articolato profilo linguistico.

4.2.1 Le caratteristiche generali del testo

A partire dal livello di segmentazione del testo in frasi e di tokenizzazione, nei corpora sono stati prima di tutto rintracciati i seguenti tratti relativi a caratteristiche formali e generali del testo:

– la lunghezza media dei periodi contenuti nei corpora, calcolata in numero di tokens,

– la lunghezza media dei tokens presenti, calcolata in caratteri.

La centralità di questo livello di monitoraggio è riconducibile ad uno dei primi suggerimenti contenuti nella "Guida per la redazione degli atti amministrativi" che invita a "formulare periodi brevi e chiari". Si è ritenuto qui pertanto interessante includerlo tra i tratti oggetto di monitoraggio. Come si può vedere dai risultati ottenuti, il confronto tra i corpora rispetto alla lunghezza media dei periodi e dei tokens contenuti permette di tratteggiare alcune loro prime caratteristiche generali.

Come mostrano i risultati riportati nelle Tabelle 4.3(a) e 4.3(b), l'intero corpus di testi giuridici (d'ora in avanti chiamato 'AMB') ha una lunghezza media sia di periodi sia di tokens maggiore rispetto sia a Rep sia a 2Par. Tuttavia, esso dimostra di avere, rispetto ad entrambi i tratti monitorati, un profilo più simile a quello di Rep che a quello di 2Par.

	Lunghezza media
AMB	26
Rep	22,26
2Par	18,67
AMBamm(Stato)	**35,70**
AMBamm(Regione)	30,24
AMBnorm(Stato)	27,16
AMBamm(Europa)	25,68
AMBnorm(Europa)	24,93
AMBnorm(Regione)	22,78
COST	16,59

(a) Lunghezza media dei periodi.

	Lunghezza media
AMB	5,60
Rep	5,06
2Par	4,98
AMBamm(Europa)	**5,79**
AMBamm(Regione)	5,66
AMBnorm(Europa)	5,62
AMBamm(Stato)	5,57
COST	5,55
AMBnorm(Regione)	5,53
AMBnorm(Stato)	5,46

(b) Lunghezza media dei tokens.

Tabella 4.3: Confronto della lunghezza media dei periodi e dei tokens nei testi normativo–amministrativi e giornalistici.

Per quanto riguarda il confronto tra le diverse tipologie di testi giuridici presi in esame, gli atti amministrativi sono i testi con la maggiore lunghezza media di periodi sia di tokens, sebbene con alcune differenze riguardo all'ente

emittente gli atti[9]. Mentre infatti la media di periodi più lunghi è degli atti statali e regionali, sono quelli comunitari e regionali ad avere la media di tokens più lunghi.

Di conseguenza, è questa la tipologia di testi giuridici ad avere caratteristiche generali più distanti da entrambi i corpora di testi giornalistici di riferimento, sebbene con una maggiore differenza rispetto a 2Par.

È inoltre interessante far osservare che la lunghezza media dei periodi di COST è la più bassa tra tutti i tipi di testi giuridici parte di AMB. La Costituzione italiana dimostra anzi di avere periodi più brevi anche di 2Par.

Come detto, se i dati relativi al monitoraggio della diversa lunghezza del periodo forniscono alcune indicazioni preliminari sul profilo dei testi in esame, i dati relativi alla lunghezza dei tokens vanno ulteriormente studiati mettendoli in rapporto con la tipologia di lessico usata nei diversi corpora.

4.2.2 Le caratteristiche morfosintattiche

Sulla base del livello di annotazione morfosintattica automatica è stato possibile monitorare i corpora rispetto alle loro caratteristiche morfosintattiche. Tali caratteristiche sono state rintracciate analizzando la diversa distribuzione delle categorie morfosintattiche presenti, con una particolare attenzione a quelle rispetto alle quali i corpora analizzati hanno dimostrato di avere le maggiori differenze.

Come si può notare dai risultati delle distribuzioni riportati nella Tabella 4.4[10], l'intero corpus giuridico differisce dai testi giornalistici soprattutto nella distribuzione di **preposizioni**, caratterizzandosi per un'occorrenza percentuale nettamente maggiore sia a Rep sia a 2Par. Inoltre, i testi giuridici mostrano di possedere una maggiore percentuale di **sostantivi** e una minore percentuale di **verbi** rispetto ai testi giornalistici di Rep. Diverso è invece il caso del confronto con 2Par, che, a differenza di Rep, dimostra di avere una percentuale di sostantivi simile a quella di AMB.

[9]Nota che nelle Tabelle 4.3(a) e 4.3(b), così come in tutte le altre tabelle contenute in questo capitolo, i sottocorpora di testi giuridici sono presentati in ordine decrescente di valore, per facilitarne il confronto.

[10]Per chiarezza nella tabella sono riportati i nomi delle categorie morfosintattiche monitorate insieme all'etichetta usata in fase di annotazione morfosintattica automatica. Per la descrizione completa vedi l'Allegato I.

Categoria morfosintattica	AMB	Rep	2Par
Aggettivi (A)	8,91	6,32	5,91
Congiunzioni (C)	4,33	4,14	3,78
Avverbi (B)	1,68	5,07	3,45
Preposizioni (E)	20,69	15,39	15,44
Determinanti (D)	0,56	0,88	1,65
Punteggiatura (F)	9,86	13,69	11,02
Interiezioni (I)	0,01	0,02	0,00
Numerali (N)	4,52	2,03	2,63
Pronomi (P)	2,02	4,28	2,24
Sostantivi (S)	30,37	26,51	29,71
Articoli (R)	6,91	8,38	10,34
Predeterminanti (T)	0,11	0,13	0,32
Verbi (V)	9,27	13,10	13,51
Residuo (X)	0,68	0,06	0,01

Tabella 4.4: Distribuzione delle categorie morfosintattiche nell'intero corpus giuridico e nei corpora giornalistici.

Rispetto a queste tre categorie morfosintattiche, come mostra la Tabella 4.5[11], gli atti amministrativi statali e regionali sono la tipologia di testo giuridico che mostra di avere *i)* la percentuale maggiore di preposizioni e di sostantivi e *ii)* la minore percentuale di verbi. In questo senso essi si differenziano di più dai corpora di testo giornalistici di riferimento.

Al contrario, COST, con la *i)* la percentuale più bassa di preposizioni e *ii)* la più alta di verbi, è il tipo di testo giuridico che più si avvicina alle distribuzioni di Rep e 2Par. Inoltre, tra questi due poli opposti rappresentati dagli atti amministrativi statali e regionali, da un lato, e dalla Costituzione italiana, dall'altro, gli atti comunitari mostrano caratteristiche morfosintattiche intermedie.

Tenendo in considerazione questi dati, l'analisi dettagliata della diversa occorrenza di **verbi**, **preposizioni** e **sostantivi** è al centro delle discussioni in quanto segue.

Infine, sebbene le oscillazioni nell'occorrenza delle **congiunzioni** non sembri rappresentare un tratto nettamente caratterizzante la lingua del diritto, tuttavia la diversa distribuzione dei diversi tipi di congiunzioni nelle varie tipologie di testi sarà ugualmente tenuta in considerazione. In questo caso,

[11]Per ragioni di spazio è riportata nella tabella solo l'etichetta usata in fase di annotazione morfosintattica automatica.

a differenza di quanto fatto per le precedenti categorie, saranno analizzate le due sottocategorie morfosintattiche previste dallo schema di annotazione: le congiunzioni coordinanti e quelle subordinanti. Un tale interesse è legato alle preliminari indicazioni sulla diversa distribuzione di principali e subordinate tra i corpora analizzati, come discusso nel Paragrafo 4.2.3.6.

	amm(S)	amm(R)	amm(E)	norm(S)	norm(R)	norm(E)	COST	Rep	2Par
A	9,02	8,41	9,79	8,18	9,27	8,95	8,73	6,32	5,91
C	4,26	3,88	4,51	4,06	3,74	4,56	5,32	4,14	3,78
B	2,06	1,70	1,89	1,27	0,98	1,69	2,15	5,07	3,45
E	21,33	21,70	20,62	21,48	21,25	19,68	18,78	15,39	15,44
D	0,59	0,48	0,77	0,38	0,27	0,73	0,72	0,88	1,65
F	10,31	9,25	8,82	10,54	11,42	9,84	8,82	13,69	11,02
I	0,01	0,00	0,00	0,00	0,03	0,00	0,00	0,02	0,00
N	4,88	5,48	2,29	6,43	5,98	3,92	2,67	2,03	2,63
P	2,23	1,98	1,83	1,98	1,74	2,14	2,25	4,28	2,24
S	29,82	31,72	29,89	30,56	31,78	29,00	29,79	26,51	29,71
R	6,21	6,03	8,03	6,13	5,57	7,81	8,58	8,38	10,34
T	0,09	0,07	0,14	0,06	0,05	0,15	0,22	0,13	0,32
V	8,73	8,74	10,36	8,13	6,62	10,51	11,79	13,10	13,51
X	0,43	0,44	0,97	0,73	1,23	0,95	0,00	0,06	0,01

Tabella 4.5: Distribuzione percentuale delle categorie morfosintattiche nei sottocorpora di testi normativo–amministrativi e nei corpora giornalistici.

4.2.2.1 Il rapporto tra sostantivi e verbi

Il tema è ampiamente dibattuto negli studi dedicati all'analisi delle differenze tra scritto e parlato così come tra varietà e generali testuali diversi. Ai fini delle discussioni qui condotte, è ricordare alcuni dati riportati da Biber (1993). Egli, analizzando come il rapporto tra sostantivi e verbi vari tra testi di prosa accademica, racconti fantastici e nel parlato, mostra come testi ad alta densità informativa come quelli accademici abbiano un rapporto più alto, contenendo una minore percentuale di verbi rispetto a testi fantastici o al parlato.

Per l'italiano, tendenza analoga è stata recentemente osservata da Montemagni (2010), dove è stato riscontrato un rapporto sostantivi/verbi più basso nei corpora di racconti fantastici e di parlato esaminati rispetto al corpus di articoli giornalistici considerato.

	Rapporto sostantivi/verbi
AMB	3,28:1
Rep	2,02:1
2Par	2,20:1
AMBnorm(Regione)	**4,80:1**
AMBnorm(Stato)	3,76:1
AMBamm(Regione)	3,63:1
AMBamm(Stato)	3,42:1
AMBamm(Europa)	2,88:1
AMBnorm(Europa)	2,76:1
COST	2,53:1

Tabella 4.6: Confronto del rapporto sostantivi/verbi nei testi normativo–amministrativi e giornalistici.

Come precedentemente fatto notare, il monitoraggio dei corpora presi in esame in questo studio ha dimostrato come l'intero corpus di testi giuridici mostri di avere una percentuale superiore di sostantivi rispetto ai testi giornalistici di confronto e una percentuale nettamente inferiore di verbi. Questi risultati suggeriscono qualche indicazione sul diverso rapporto di queste due categorie morfosintattiche nelle due tipologie di testi.

Il dato è riportato nella Tabella 4.6, dove sono messi a confronto i risultati del rapporto tra la distribuzione di sostantivi e verbi nell'intero corpus giuridico, in Rep e 2Par e nei diversi tipi di testi giuridici.

Come ci si poteva aspettare, AMB ha un rapporto superiore a quello rintracciato nei due corpora giornalistici, dimostrando in particolari valori più simili a quelli di 2Par. Come fatto precedentemente osservare, infatti, 2Par contiene una percentuale superiore rispetto a Rep di sostantivi, avvicinandosi così maggiormente alle distribuzioni di AMB.

Inoltre, coerentemente con quanto osservato prima a proposito dell'elevata percentuale di occorrenza dei sostantivi, gli atti amministrativi e normativi statali e regionali sono la tipologia di testi giuridico con il rapporto più alto. Il rapporto più basso si ha invece nella Costituzione, testo giuridico nel quale era risultato esserci la percentuale maggiore di verbi. Infine, come notato prima, gli atti comunitari hanno un comportamento intermedio tra questi due poli opposti.

Queste osservazioni suggeriscono pertanto una caratterizzazione dei testi giuridici, e in particolare degli atti amministrativi statali e regionali, come

testi altamente informativi, contraddistinti da una bassa percentuale di verbi e da un'elevata occorrenza di sostantivi.

4.2.2.2 La distribuzione dei verbi

Il monitoraggio della distribuzione dei verbi si è concentrato sui due seguenti aspetti.

I modi verbali

Facendo riferimento ai tratti morfologici individuati in fase di annotazione morfosintattica, è stato possibile monitorare la diversa distribuzione, ad esempio, dei modi verbali. Tra tutti i risultati ottenuti, è qui di particolare interesse riportare i dati relativi alla distribuzione delle **forme participiali**.

Come messo infatti in luce da Garavelli (2001, p. 162), infatti, una delle caratteristiche dei testi giuridici è l'uso massiccio "frasi ridotte participiali", in linea con una spiccata propensione alla "sintesi strutturale". Al contrario, la "Guida per la redazione degli atti amministrativi" invita espressamente ad "evitare i costrutti sintetici come [...] le forme implicite del verbo, come gerundi o participi, quando potrebbero essere usate le forme esplicite".

È necessario qui chiarire che i dati raccolti rispetto a questo tratto sono sovrastimati, dal momento che i modi verbali sono stati calcolati token per token. Ciò implica che un tempo composto come *è stato adottato* non è stato considerato come un tutt'uno, ma il verbo *essere* e il verbo *adottare* sono stati considerati due singole occorrenze di verbi con modo participio. Tuttavia, poiché la stessa metodologia è stata adottata in tutti i corpora monitorati, le distribuzioni osservate si possono considerare significative e affidabili.

Come mostra infatti la Figura 4.1, l'intero corpus di testi giuridici ha una percentuale nettamente maggioritaria di forme participiali (pari al 37,51% di tutti i modi presenti) rispetto a Rep (12,67%) e 2Par (5,12%).

Inoltre, gli atti statali e regionali sia normativi sia amministrativi si differenziano dalle altre tipologie di atti per una maggiore distribuzione percentuale di participi; mentre la Costituzione con il 22,31% di forme participiali è il testo che più si avvicina ai testi giornalistici di riferimento.

I dati ottenuti forniscono dunque una conferma quantitativa a quanto osservato da Garavelli (2001).

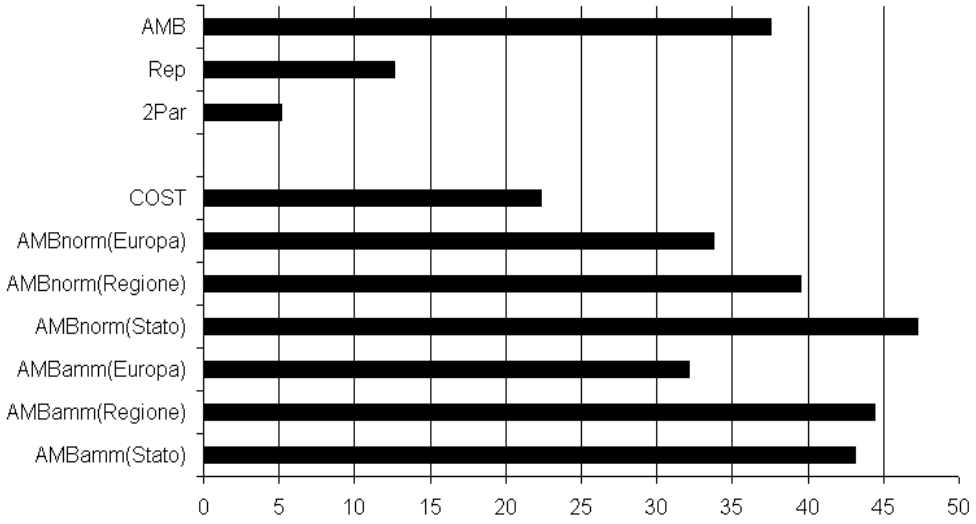


Figura 4.1: La distribuzione percentuale di forme participiali nei testi normativo–amministrativi e giornalistici.

Le persone del verbo

Sempre a partire dall'annotazione dei tratti morfologici, è stato possibile studiare anche l'uso delle persone del verbo. Anche in questo caso i risultati (riportati nella Tabella 4.7) sono in linea con quanto osservato da Garavelli (2001, p. 118), dove la bassa occorrenza della I persona singolare e della II plurale è annoverata tra le più vistose "assenze" notate nei testi giuridici esaminati.

Una loro inferiore occorrenza percentuale rispetto a quella riscontrata nei testi giornalistici è infatti una evidente caratteristica rintracciata nell'intero corpus di testi normativo–amministrativi. Tuttavia, va fatto notare che 2Par si discosta parzialmente da questa tendenza mostrando una bassissima percentuale di forme verbali alla II persona plurale.

Altri due dati sono chiaramente distinguibili dall'analisi delle differenze: *i)* la quasi totale assenza di forme di I persona plurale nei testi giuridici, contrariamente a quanto avviene nei testi giornalistici, e *ii)* la netta preponderanza di forme di III persona plurale in AMB rispetto a Rep. In quest'ultimo caso, 2Par mostra invece una distribuzione simile a quella di AMB.

Rispetto a questi dati, gli atti amministrativi e normativi regionali dimo-

	I sing	II sing	III sing	I plur	II plur	III plur
AMB	1,93	0,78	59,12	0,00	0,13	44,19
Rep	5,51	0,91	68,67	3,53	0,46	20,92
2Par	4,30	0,27	45,41	5,32	0,06	44,65
AMBamm(Stato)	1,20	0,81	55,53	0,00	0,23	40,59
AMBamm(Regione)	3,22	1,39	64,87	0,00	0,27	30,24
AMBamm(Europa)	1,98	0,79	60,32	0,00	0,00	81,94
AMBnorm(Stato)	1,98	0,68	58,39	0,00	0,12	38,82
AMBnorm(Regione)	3,39	0,00	55,53	0,00	0,23	40,59
AMBnorm(Europa)	0,58	1,03	57,33	0,01	0,09	40,97
COST	1,18	0,79	61,89	0,00	0,00	36,15

Tabella 4.7: Distribuzione percentuale delle forme della persona verbale nei testi normativo–amministrativi e giornalistici.

strano di essere la tipologia di atti che con la percentuale più alta di forme verbali alla I persona singolare e alla II plurale si allontanano di più dalla tendenza riscontrata in tutto AMB. Inoltre, gli atti amministrativi comunitari, seguiti da quelli normativi comunitari, sono i testi giuridici con la percentuale maggiore di III persone plurali. L'intuizione è che ciò sia legato ai destinatari degli atti comunitari, gli stati membri della Comunità europea.

4.2.2.3 La distribuzione delle preposizioni

Nel suo studio Biber (1993) mette in stretta relazione tra tratti rintracciabili in un corpus: l'alta percentuale di occorrenza di nomi, di complementi preposizionali e di aggettivi attributivi. In base ai dati da lui raccolti, tali tratti infatti cooccorrendo in un corpus costituiscono una significativa dimensioni di variazione tra varietà e registri testuali.

Sulla scia di questa osservazione, è qui interessante far notare che la distribuzione di preposizioni nelle diverse tipologie di corpora esaminati segue i dati relativi al rapporto sostantivi/verbi. Come mostra infatti la Figura 4.2, AMB caratterizzato da una percentuale maggiore di sostantivi e una minore di verbi si differenzia sia da Rep sia da 2Par per una percentuale maggiore di preposizioni.

Allo stesso modo, gli atti amministrativi e normativi statali e regionali che avevano dimostrato di avere un rapporto sostativi/verbi più elevato dei corrispondenti atti comunitari e della Costituzione sono la tipologia di testo giuridico con la percentuale maggiore di preposizioni.

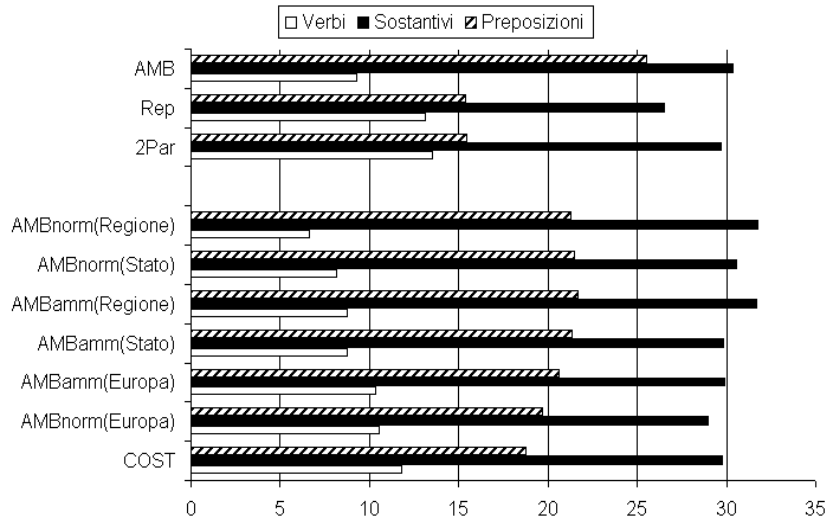


Figura 4.2: La distribuzione percentuale di sostantivi, verbi e preposizioni nei testi normativo–amministrativi e giornalistici.

Si è ritenuto qui importante sottolineare questo dato dal momento che esso fornisce alcune preliminari indicazioni su un comportamento sintattico (discusso nei paragrafi successivi) nettamente distintivo dei testi normativo–amministrativi rispetto a quelli giornalistici, quello relativo cioè alle lunghe 'catene' di complementi preposizionali modificatori di sostantivi.

4.2.2.4 Il rapporto tra congiunzioni coordinanti e subordinanti

Si è deciso di monitorare i corpora in esame anche rispetto a questo tratto dal momento che i risultati ottenuti forniscono preliminari indizi di una tendenza riscontrata in fase di analisi delle caratteristiche sintattiche: la minore frequenza di strutture ipotattiche rispetto alle distribuzioni riscontrate nei testi giornalistici di "La Repubblica".

Sebbene i dati riportati in questo paragrafo non possano esserne considerati una causa diretta, tuttavia è interessante far notare come dalla Figura 4.3 risulti chiaramente che, rispetto a questo tratto monitorato, l'intero corpus giuridico ha un comportamento diverso soprattutto da Rep. Rispetto al 29,01% di congiunzioni subordinanti presenti in Rep, questa sottocatego-

ria morfosintattica rappresenta il 13,64% sul totale di tutte le occorrenze di congiunzioni in AMB[12].

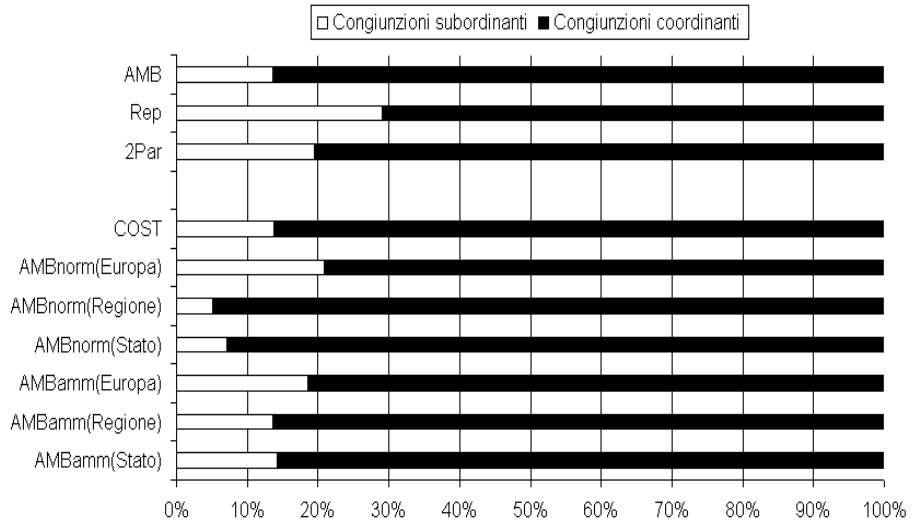


Figura 4.3: La distribuzione percentuale di congiunzioni coordinanti e subordinanti nei testi normativo–amministrativi e giornalistici.

I testi giuridici si differenziano dunque da quelli giornalistici per il diverso rapporto tra i due tipi di congiunzioni. Come mostrano infatti i dati riportati nella Tabella 4.8, AMB è caratterizzato da un rapporto più alto rispetto sia a Rep sia 2Par, nonostante quest'ultimo abbia un rapporto superiore a Rep.

Inoltre, tra le diverse tipologie di testi giuridici, gli atti con la minore frequenza di congiunzioni subordinanti sono quelli statali e regionali, e in particolare quelli normativi. Questi ultimi due tipi sono infatti i testi con il rapporto congiunzioni coordinanti/subordinanti più alto. Al contrario, gli atti comunitari normativi e amministrativi sono la tipologia di testo giuridico con la maggiore frequenza di congiunzioni subordinanti e, di conseguenza, con i più bassi valori del rapporto tra i due tipi di congiunzioni.

[12]È necessario qui ricordare, come già fatto notare, che il monitoraggio di questo tratto si basa sull'annotazione automatica delle sottocategorie morfosintattiche previste dallo schema di annotazione riportato nell'Allegato I.

	Rapporto
AMB	6,33:1
Rep	2,45:1
2Par	4,14:1
AMBnorm(Regione)	**18,68:1**
AMBnorm(Stato)	12,92:1
AMBamm(Regione)	6,33:1
COST	6,25:1
AMBamm(Stato)	6,05:1
AMBamm(Europa)	4,40:1
AMBnorm(Europa)	3,79:1

Tabella 4.8: Confronto del rapporto congiunzioni coordinanti/subordinanti sul totale di congiunzioni presenti nei testi normativo–amministrativi e giornalistici.

4.2.3 Le caratteristiche sintattiche

Sulla base del livello di annotazione sintattica a dipendenze è stato possibile monitorare i corpora rispetto alle loro caratteristiche sintattiche.

È importante ricordare qui che, sebbene questo livello di annotazione linguistica automatica sia in generale il meno affidabile e in particolare per i testi giuridici[13], tuttavia un'attenta scelta dei risultati ottenuti ha permesso di rintracciare nei corpora in esame alcuni tratti significativi per la ricostruzione del loro profilo sintattico.

4.2.3.1 La distribuzione delle relazioni di dipendenza

Come nel caso dell'analisi della distribuzione delle categorie morfosintattiche, anche in questo caso il confronto tra i vari corpora rispetto alla diversa distribuzione dei tipi di relazione di dipendenza annotati in modo automatico ha permesso solo alcune preliminari riflessioni, che sono poi state approfondite grazie alle successive analisi condotte intrecciando i dati ottenuti da questo livello di annotazione linguistica.

I risultati del monitoraggio comparativo *i)* tra l'intero corpus giuridico e i corpora giornalistici e *ii)* tra le diverse tipologie di testi giuridici sono riportati rispettivamente nella Tabelle 4.9 e 4.10.

[13]Per questo aspetto vedi le discussioni del Capitolo 3.

Relazione di dipendenza	AMB	Rep	2Par
comp_loc	0,41	0,86	1,37
clit	0,35	0,91	0,49
con	4,26	4,06	4,05
mod_temp	0,05	0,43	0,42
arg	1,15	1,94	1,62
disj	0,46	0,08	0,18
subj	2,41	4,59	5,64
conj	3,81	3,47	3,99
subj_pass	0,44	0,22	0,13
sub	0,44	1,10	0,66
pred	0,61	1,57	1,59
comp_ind	0,03	0,21	0,09
concat	0,01	0,04	0,02
aux	1,05	1,91	2,20
ROOT	6,27	5,98	5,78
prep	**20,54**	15,29	15,44
comp_temp	0,21	0,36	0,78
comp	**17,41**	12,30	11,13
obj	2,63	3,59	4,78
mod	19,83	17,14	16,79
punc	8,39	12,33	9,69
mod_loc	0,00	0,10	0,05
det	6,89	8,37	10,34
modal	0,63	0,58	0,81
neg	0,41	0,87	0,36
mod_rel	0,56	1,43	1,29
dis	0,56	0,14	0,28

Tabella 4.9: Distribuzione delle relazioni di dipendenza nell'intero corpus giuridico e nei corpora giornalistici.

Il dato che risulta subito evidente analizzando le differenze tra AMB, da un lato, e Rep e 2Par, dall'altro, è la netta preponderanza nel corpus giuridico di relazioni di tipo 'comp' e 'prep'[14].

[14]Le relazioni di dipendenza a cui si fa riferimento qui e nei successivi paragrafi sono quelle parte dello schema di annotazione descritto nell'Allegato I. Per chiarezza di lettura, in questo capitolo si riporta di volta in volta la definizione di quelle oggetto di monitoraggio. Pertanto, la relazione 'complement' ('comp') è la relazione tra una testa e un complemento preposizionale, sia esso modificatore o argomento. Questa relazione funzionale sottospecificata è particolarmente utile in quei casi in cui è difficile stabilire la natura

	amm(S)	amm(R)	amm(E)	norm(S)	norm(R)	norm(E)	COST	Rep	2Par
comp_loc	0,32	0,45	0,49	0,34	0,41	0,44	0,38	0,86	1,37
clit	0,45	0,37	0,34	0,26	0,23	0,31	0,49	0,91	0,49
con	4,40	3,95	3,77	4,40	4,82	3,81	4,66	4,06	4,05
mod_temp	0,04	0,04	0,07	0,04	0,02	0,08	0,08	0,43	0,42
arg	1,23	1,26	1,64	0,98	0,54	1,47	0,94	1,94	1,62
disj	0,31	0,28	0,58	0,39	0,22	0,61	0,79	0,08	0,18
subj	2,08	1,81	2,89	1,86	1,69	2,92	3,64	4,59	5,64
conj	3,81	3,59	3,28	3,92	4,52	3,39	4,16	3,47	3,99
subj_pass	0,31	0,35	0,44	0,40	0,42	0,48	0,70	0,22	0,13
sub	0,44	0,38	0,63	0,22	0,11	0,67	0,65	1,10	0,66
pred	0,59	0,45	0,77	0,40	0,36	0,72	0,97	1,57	1,59
comp_ind	0,01	0,02	0,03	0,01	0,01	0,03	0,08	0,21	0,09
concat	0,01	0,02	0,01	0,01	0,00	0,01	0,00	0,04	0,02
aux	0,84	0,87	0,82	0,99	0,89	1,16	1,76	1,91	2,20
ROOT	5,20	5,70	6,23	6,16	6,86	6,76	6,99	5,98	5,78
prep	**21,14**	**21,49**	20,47	**21,31**	**21,11**	19,52	18,73	15,29	15,44
comp_temp	0,30	0,22	0,20	0,28	0,12	0,22	0,15	0,36	0,78
comp	**18,08**	**18,22**	16,98	**18,02**	**18,50**	16,22	15,84	12,30	11,13
obj	2,36	2,55	3,28	2,16	1,67	3,09	3,34	3,59	4,78
mod	21,03	22,20	18,54	20,79	21,12	19,24	15,91	17,14	16,79
punc	8,74	7,90	7,54	9,00	9,56	8,33	7,67	12,23	9,69
mod_loc	0,01	0,00	0,00	0,00	0,00	0,01	0,00	0,10	0,05
det	6,20	6,02	8,01	6,12	5,57	7,78	8,50	8,37	10,34
modal	0,59	0,45	1,00	0,46	0,24	0,71	0,99	0,58	0,81
neg	0,37	0,26	0,42	0,30	0,19	0,39	0,93	0,87	0,36
mod_rel	0,64	0,54	0,57	0,51	0,41	0,63	0,61	1,43	1,29
dis	0,37	0,37	0,75	0,51	0,29	0,79	0,87	0,14	0,28

Tabella 4.10: Distribuzione percentuale delle relazioni di dipendenza nei testi normativo–amministrativi e giornalistici.

In particolare, tra le varie tipologie di testi giuridici, gli atti amministrativi e normativi statali e regionali spiccano tra tutti per questa tendenza; mentre, la Costituzione è il tipo di testo che più si allontana da una tale distribuzione.

Il dato è strettamente legato a quello relativo alla distribuzione delle

argomentale o di modificatore del complemento.

La relazione 'preposition' ('prep') è la relazione tra una testa preposizionale e il suo complemento, sia esso frasale o meno.

categorie morfosintattiche. Nel Paragrafo 4.2.2.3, era stato fatto notare come AMB si differenziasse dai corpora giornalistici di riferimento per una spiccata occorrenza di preposizioni e come, in particolare, gli atti amministrativi e normativi statali e regionali si contraddistinguessero tra tutti i testi giuridici per questa tendenza.

In base allo schema di annotazione sintattica, i tokens di tipo preposizionale costituiscono la testa della relazione 'prep' e il dipendente della relazione 'comp', come illustrato nella Figura 4.4.

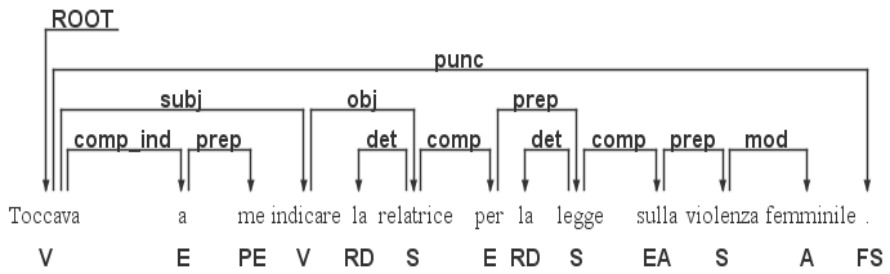


Figura 4.4: Esempio di come i tokens di tipo preposizionale (E o EA) costituiscono la testa della relazione 'prep' e il dipendente della relazione 'comp'.

Considerato dunque il fatto che in tutto il corpus di testi giuridici le preposizioni sono un quinto in più di quelle nei corpora di testi giornalistici, è questo il motivo per cui anche le relazioni 'comp' e 'prep' sono i due tipi di relazione di dipendenza rispetto ai quali AMB si differenzia di più sia da Rep sia da 2Par.

Inoltre, la diversa distribuzione di preposizioni, sostantivi e verbi discussa nel Paragrafo 4.2.2.3 ha ripercussioni sui modi della coordinazione. In base allo schema di annotazione sintattico adottato, i modi della coordinazione sono rintracciabili sulla base della distribuzione delle relazioni 'con' (e 'dis') e 'conj' ('disj')[15].

[15]La relazione 'conjunct linked by a copulative conjunction' ('conj') è la relazione che unisce il secondo (o il terzo, quarto, ecc..) elemento parte di una struttura coordinata al primo token, il quale rappresenta la testa sintattica dell'intera struttura; sempre usato in coppia con la relazione 'con'. La relazione 'conjunct in a disjunctive compound linked by a disjunctive conjunction' ('disj') è la relazione che il secondo (o il terzo, quarto, ecc..)

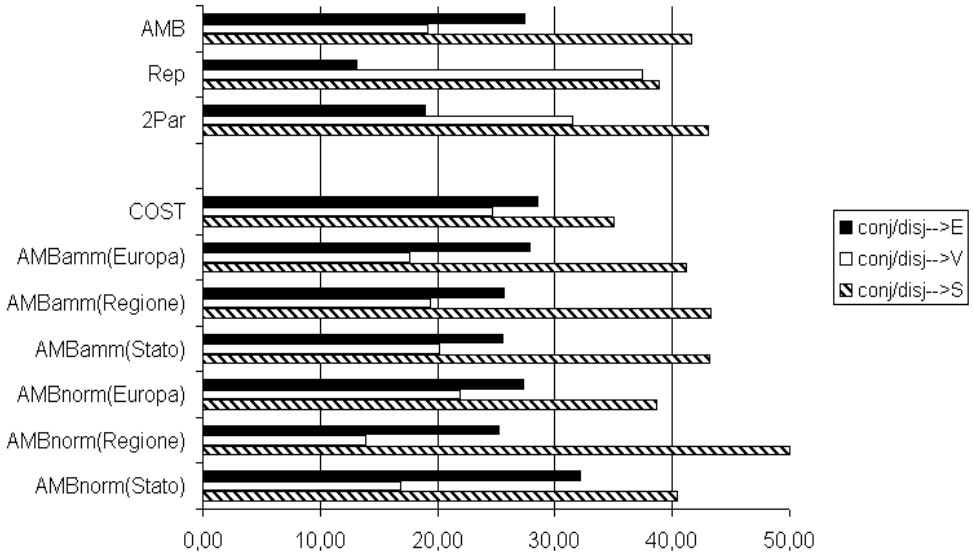


Figura 4.5: Classificazione comparativa nei testi normativo–amministrativi e giornalistici delle costruzioni coordinate (o disgiunte) sulla base della categoria morfosintattica dei tokens che coordinano.

Sono qui ora oggetto di attenzione le relazioni 'conj' e 'disj', dal momento che permettono di mettere in luce una peculiarità dei testi giuridici. Come mostrato infatti nella Figura 4.5, AMB si differenzia sia da Rep sia da 2Par per una maggiore frequenza di relazioni di tipo 'conj' e 'disj' che coordinano tokens di tipo preposizionale. Al contrario, nettamente meno frequenti sono in AMB le relazioni che coordinano verbi.

In particolare, gli atti normativi statali e regionali, parte della tipologia di atti con la maggiore frequenza percentuale di preposizione e la minore di verbi, sono quelli con *i)* la maggiore occorrenza di relazioni che coordinano o disgiungono tokens preposizionali e *ii)* la minore percentuale di relazioni che legano verbi. Mentre, la Costituzione, caratterizzata da un comportamento morfosintattico opposto, contiene la più alta percentuale di relazioni 'conj' e 'disj' che legano verbi.

---

elemento parte di una struttura coordinata al primo token, il quale rappresenta la testa sintattica dell'intera struttura; sempre usato in coppia con la relazione 'dis'.

4.2.3.2 La lunghezza delle relazioni di dipendenza

Il monitoraggio di questo tratto, insieme a quello discusso nel Paragrafo 4.2.3.3, permette di mettere a confronto i corpora analizzati rispetto alla struttura degli alberi sintattici dei loro periodi. La misura della lunghezza delle relazioni di dipendenza che strutturano sintatticamente un periodo e il modo in cui esse sono organizzate in maniera gerarchica all'interno dell'albero sintattico del periodo permettono infatti di indagare più nel dettaglio quali sono le caratteristiche sintattiche distintive dei vari corpora presi in esame.

Sulla base dell'annotazione sintattica automatica, la lunghezza delle relazioni di dipendenza viene calcolata come la distanza tra una testa sintattica e il suo dipendente legati da una relazione di dipendenza. È in particolare monitorata la lunghezza media della distanza **massima** tra due tokens parte di una coppia testa–dipendente. Operativamente, la lunghezza viene calcolata sulla base del numero di tokens che intercorrono tra i due elementi della coppia.

Ad esempio, nel seguente periodo la relazione di dipendenza più lunga è quella di 'subj'[16] che lega il token *attività* alla sua testa sintattica *sono*:

- *Le **attività** di trasporto e dispacciamento del gas naturale a rete, nonché la gestione di infrastrutture di approvvigionamento di energia connesse alle attività di trasporto e dispacciamento di energia a rete, **sono** di interesse pubblico e sono sottoposte agli obblighi di servizio pubblico derivanti dalla normativa comunitaria, dalla legislazione vigente e da apposite convenzioni con le autorità competenti.*

Come si può vedere nell'estratto di annotazione riportato nella Figura 4.6, tra la testa sintattica e il suo dipendente c'è una distanza di 32 tokens (compresa la punteggiatura).

I risultati del monitoraggio, riportati nella Tabella 4.11, dimostrano che la lunghezza media delle relazioni nell'intero corpus giuridico è nettamente maggiore rispetto a quella in Rep e, soprattutto, a quella in 2Par.

In particolare, gli atti amministrativi statali e regionali sono la tipologia di testo giuridico ad avere le relazioni più lunghe. Mentre, la Costituzione è il corpus con le relazioni più brevi, con valori inferiori anche a quelli di 2Par.

[16] La relazione 'subject' ('subj') è la relazione che lega un verbo attivo al suo soggetto.

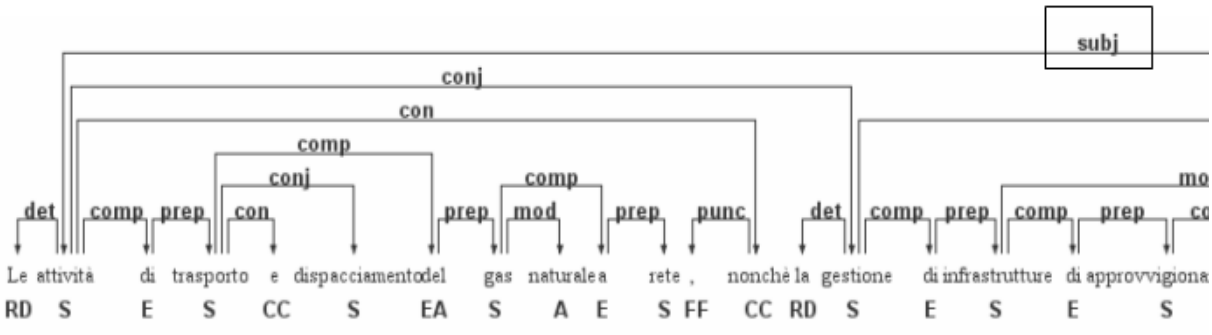


Figura 4.6: Un estratto della relazione di dipendenza 'subj' lunga 32 tokens che lega il dipendente *attività* alla sua testa sintattica *sono*.

	Lunghezza media
AMB	13,62
Rep	8,80
2Par	7,71
AMBamm(Stato)	**19,08**
AMBamm(Regione)	15,39
AMBnorm(Stato)	14,82
AMBnorm(Europa)	12,50
AMBnorm(Regione)	12,39
AMBamm(Europa)	12,35
COST	7,21

Tabella 4.11: Confronto della lunghezza media delle relazioni di dipendenza massime nei testi normativo–amministrativi e giornalistici.

4.2.3.3 Il livello di incassamento gerarchico

Come anticipato nel paragrafo precedente, anche il monitoraggio di questo tratto permette di ricostruire le caratteristiche dei corpora analizzati per quanto riguarda la struttura degli alberi sintattici dei periodi in essi contenuti.

In questo caso, il modo in cui le relazioni di dipendenza si organizzano in maniera gerarchica all'interno dell'albero sintattico di un periodo è stato qui monitorato analizzando l'altezza **massima** dell'albero sintattico, operativamente calcolata come il numero di relazioni di dipendenza consecutive ('a

cascata') tra una foglia (rappresentata da tokens del testo senza dipendenti) e la radice dell'albero.

Ad esempio, nel seguente periodo, la massima distanza che intercorre tra una foglia e la radice dell'albero sintattico è quella tra il token *eseguire* e la radice *hanno*:

- *I proprietari, possessori o detentori a qualsiasi titolo dei beni indicati al comma 1,* ***hanno*** *l'obbligo di sottoporre alla Regione i progetti delle opere di qualunque genere che intendano* ***eseguire****, al fine di ottenere la preventiva autorizzazione.*

Come si può vedere nella Figura 4.7, si tratta di una serie di 8 dipendenze 'a cascata' (numerate all'interno della cornice tratteggiata) di tipo (nell'ordine) 'obj', 'arg', 'prep', 'obj', 'comp', 'prep', 'mod_rel' e 'arg'.

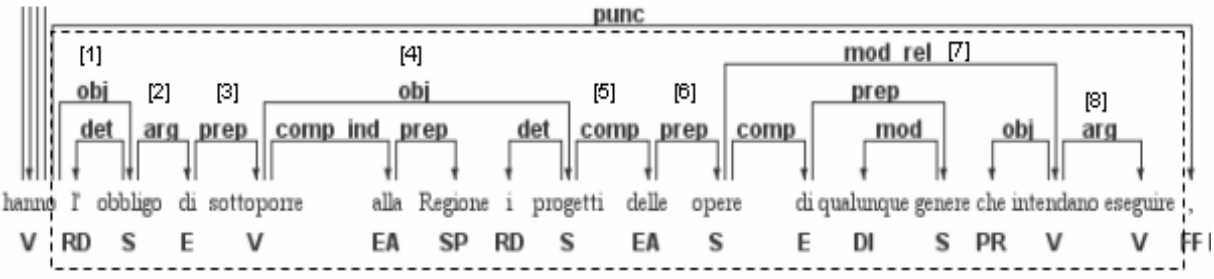


Figura 4.7: Un esempio di periodo con 8 dipendenze 'a cascata'.

Come dimostrano i dati riportati nella Tabella 4.12, il monitoraggio dei corpora rispetto a questo tratto ha messo in luce che *i)* l'intero corpus giuridico ha una media di altezze massime degli alberi sintattici dei periodi in esso contenuti superiore a quella degli alberi dei periodi di Rep e 2Par, che *ii)* gli atti amministrativi sono la tipologia di testo giuridico con periodi caratterizzati dagli alberi sintattici più alti di tutto AMB, che *iii)* al contrario, la Costituzione contiene periodi con gli alberi sintattici più bassi e che *iv)* in generale gli atti statali e regionali hanno periodi con alberi più alti dei corrispettivi (per tipologia di testo) atti comunitari.

	Altezza media
AMB	6,23
Rep	5,71
2Par	5,26
AMBamm(Stato)	**7,76**
AMBamm(Regione)	7,20
AMBamm(Europa)	6,45
AMBnorm(Stato)	6,24
AMBnorm(Europa)	5,96
AMBnorm(Regione)	5,40
COST	4,58

Tabella 4.12: Confronto dei valori relativi alla media dell'altezza massima degli alberi sintattici nei testi normativo–amministrativi e giornalistici.

4.2.3.4 Le dipendenze di predicati verbali

Sempre a partire dal livello di annotazione sintattica a dipendenze, sono state analizzate le proprietà distribuzionali dei predicati verbali presenti nei corpora in esame. Lo studio di questo tratto è stato condotto sulla base dei dipendenti direttamente governati da una testa verbale, di qualsiasi natura essi siano. Sebbene nel monitoraggio di questo tratto non sia stata fatta distinzione tra dipendenti di tipo nominale, argomenti sottocategorizzati dal verbo e modificatori di varia natura (locativi, temporali, causali, ecc...), è stato tuttavia possibile raccogliere alcuni dati significativi circa la struttura valenziale dei verbi presenti nei testi in esame. Al momento, infatti, una distinzione automatica tra i due tipi di dipendenze condotta sulla base dell'annotazione sintattica automatica non è stata ritenuta affidabile.

I risultati del monitoraggio, mostrati nella Tabella 4.13, hanno permesso di mettere in luce come l'intero corpus giuridico contiene predicati verbali caratterizzati da un numero medio di dipendenti inferiore rispetto a Rep e 2Par, sebbene la differenza con i valori riscontrati in Rep non possa essere considerata statisticamente significativa.

Inoltre, tra tutte le tipologie di testi giuridici esaminati, la Costituzione risulta il corpus con il numero medio di dipendenti più alto, seguita (in ordine descrescente) dagli atti regionali, statali e comunitari, con l'unica eccezione degli atti amministrativi regionali che mostrano i valori più bassi.

Questo dato diventa ancora più significativo se ulteriormente indagato rispetto alla distribuzione percentuale delle teste verbali per numero di di-

	Numero medio di dipendenti
AMB	2,08
Rep	2,11
2Par	2,21
COST	**2,27**
AMBnorm(Regione)	2,15
AMBamm(Stato)	2,08
AMBnorm(Stato)	2,08
AMBnorm(Europa)	2,03
AMBamm(Europa)	2,01
AMBamm(Regione)	1,96

Tabella 4.13: Confronto del numero medio dei dipendenti di una testa verbale per periodo nei testi normativo–amministrativi e giornalistici.

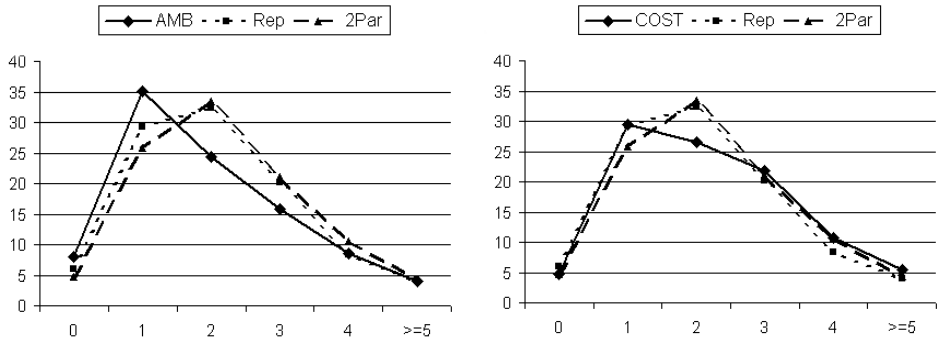
pendenti. Come mostra la Figura 4.8, parte (a), l'intero corpus giuridico (AMB) è caratterizzato *i)* da una frequenza maggiore di verbi con un solo dipendente rispetto a Rep e 2Par e *ii)* da una frequenza minore di verbi con due e più dipendenti. Nelle parti (b), (c) e (d) della figura è inoltre visualizzata la distribuzione delle teste verbali nella Costituzione italiana e nei corpora di atti amministrativi e normativi.

Le ragioni di questi andamenti necessiterebbero di analisi più approfondite che al momento non è stato possibile realizzare. Ovviamente, dipendono prima di tutto dalle scelte lessicali, diverse nelle due tipologie di testi.

L'intuizione è che essi possano essere inoltre legati all'occorrenza di forme participiali, nettamente maggioritarie nei testi giuridici che in quelli giornalistici. Come precedentemente osservato, un tale dato può essere legato ad una maggiore frequenza di frasi participiali, che non richiedono un soggetto esplicito, o alla presenza di forme verbali passive. Una tale ipotesi è discussa nel Capitolo 7, dove è connessa con questioni di rappresentazione di materiale semantico non linguisticamente realizzato nel testo ma fondamentale per la descrizione delle proprietà semantico–combinatorie di verbi.

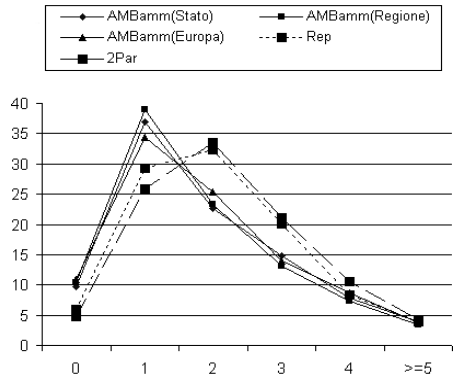
4.2.3.5 Le forme della modificazione nominale

È stato inoltre considerato interessante monitorare i corpora in esame rispetto ai modificatori di teste nominali, con un particolare riguardo al numero medio di complementi preposizionali dipendenti in sequenza ('a cascata') da una

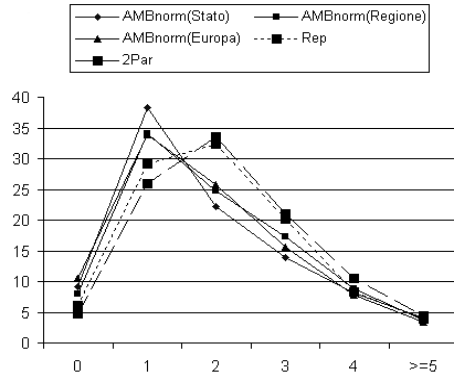


(a) Intero corpus giuridico.

(b) Costituzione.



(c) Atti amministrativi.



(d) Atti normativi.

Figura 4.8: Confronto della distribuzione percentuale delle teste verbali per numero di dipendenti nei testi normativo–amministrativi e giornalistici.

testa nominale e al livello di incassamento gerarchico con cui i complementi si distribuiscono nel periodo.

L'attenzione per il monitoraggio di questo tratto linguistico trova una giustificazione nello studio di Garavelli (2001), dove viene fatto notare che la propensione per l'uso di sostantivi per lo più astratti fa sì che siano "specialmente i nessi, i grappoli di astrazioni concatenate in 'complementi del nome' a marcare sintatticamente (e testualmente) gli enunciati". La conseguenza più significativa di tali "complicazioni strutturali" è che esse possono

diventare "fonte di oscurità o di difficoltà interpretative"[17].

Operativamente, questo tratto è stato monitorato calcolando la distribuzione di sequenze consecutive di relazioni di dipendenza di tipo 'prep'[18]. Ad esempio, come mostra la rappresentazione grafica della struttura sintattica del seguendo periodo riportata nella Figura 4.9, il sostantivo *accordo* è testa nominale di una sequenza 'a cascata' di 7 complementi preposizionali (segnalata dalla cornice tratteggiata):

- *Il Consiglio è giunto ad un* ***accordo sui*** *contributi* ***dei*** *singoli Stati membri* ***all****'adempimento* ***dell****'impegno globale* ***di*** *riduzione* ***delle*** *emissioni* ***della*** *Comunità nelle conclusioni del Consiglio del 16 giugno 1998.*

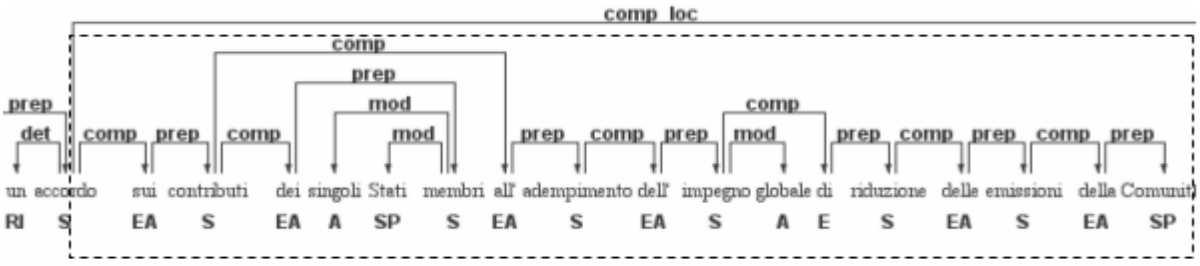


Figura 4.9: Un esempio di frase con una sequenza di 7 complementi preposizionali 'a cascata'.

È dunque allo scopo di trovare una conferma quantitativa di quanto affermato da Bice Mortara Garavelli che i corpora in esame sono stati monitorati rispetto *i)* alla profondità media delle catene di modificatori nominali di tipo preposizionale che occorrono in un periodo e *ii)* alla distribuzione degli incassamenti gerarchici per livello di profondità.

[17]Vedi Garavelli (2001, pp. 171–175).

[18]Per chiarezza si ricorda qui che, sulla base dello schema di annotazione sintattica a dipendenze adottato, la relazione 'preposition' ('prep') è la relazione tra una testa preposizionale e il suo complemento, sia esso frasale o meno.

	Profondità media
AMB	1,81
Rep	1,43
2Par	1,34
AMBamm(Regione)	**1,94**
AMBnorm(Regione)	1,89
AMBamm(Stato)	1,88
AMBnorm(Stato)	1,86
AMBnorm(Europa)	1,77
AMBamm(Europa)	1,76
COST	1,53

Tabella 4.14: Confronto dei valori relativi alla profondità media delle sequenze di complementi preposizionali, gerarchicamente organizzati, nei testi normativo–amministrativi e giornalistici.

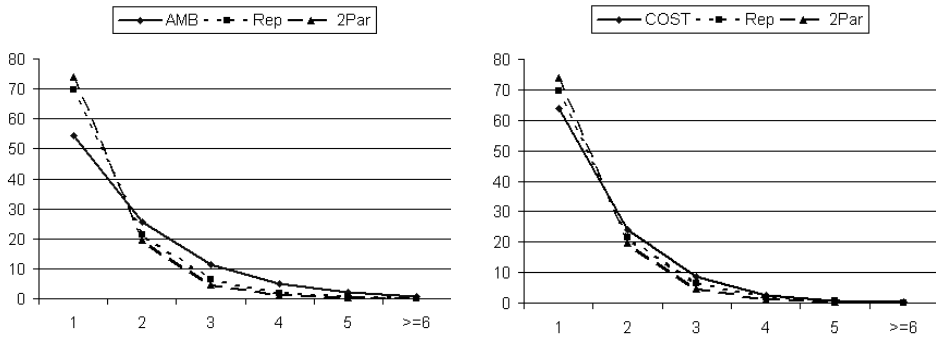
La profondità media delle 'catene' di complementi preposizionali

Come si può vedere nella Tabella 4.14, l'intero corpus di testi giuridici mostra di contenere sequenze di complementi preposizionali modificatori di teste nominali più profondi di quelle presenti in Rep e in particolare in 2Par.

In particolare, gli atti regionali amministrativi e normativi dimostrano di essere la tipologia di testo con le sequenze più lunghe, seguiti poi da quelli statali (nell'ordine, amministrativi e normativi) e dagli atti comunitari; mentre la Costituzione risulta essere il testo giuridico con le 'catene' di complementi preposizionali più brevi.

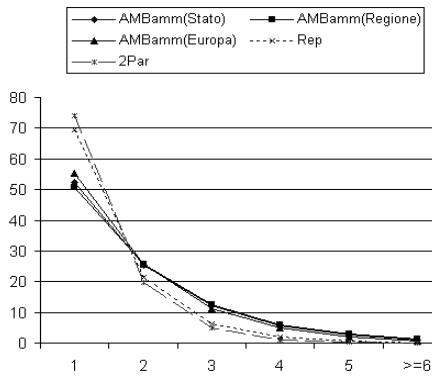
La distribuzione delle 'catene' di complementi preposizionali per livello di profondità

I risultati si qui ottenuti sono ancora più significativi se intrecciati con quelli ottenuti dal monitoraggio dei diversi corpora rispetto alla distribuzione percentuale delle 'catene' di complementi preposizionali per livello di profondità. Come mostra la Figura 4.10, parte (a), l'intero corpus giuridico (AMB) ha *i)* una percentuale inferiore a Rep e 2Par di sequenze lunghe 1 e *ii)* una percentuale superiore di sequenze lunghe più di 2. Sebbene poi la figura non permetta di visualizzare il dato, AMB contiene sequenze lunghe fino a 8 complementi incatenati, anche se con frequenze di occorrenza molto basse; mentre, Rep e 2Par si fermano invece a sequenze di 5 complementi.

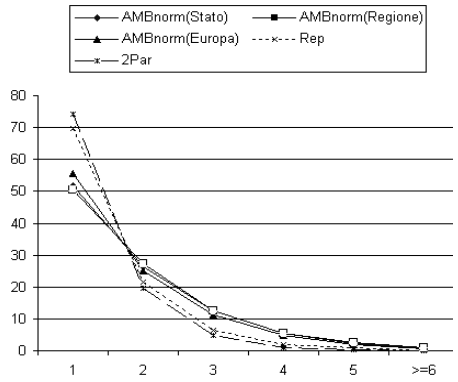


(a) Intero corpus giuridico.

(b) Costituzione.



(c) Atti amministrativi.



(d) Atti normativi.

Figura 4.10: Confronto della distribuzione percentuale delle catene di complementi preposizionali per livello di profondità nei testi normativo–amministrativi e giornalistici.

Inoltre, le singole parti della figura mostrano come nella Costituzione, parte (b), la differenza di distribuzione delle catene di lunghezza pari a 1 e ≥ 2 rispetto ai corpora giornalistici diminuisce e come gli atti comunitari sia amministrativi, parte (c), e normativi, parte (d), abbiano un comportamento più simile a Rep e 2Par rispetto ai corrispettivi atti regionali e statali. Sebbene, anche in questo caso la figura non permetta di visualizzarlo, è interessante far notare che mentre gli atti statali e regionali contengono sequenze

lunghe fino a 11 o 12 complementi preposizionali, anche se con frequenze di occorrenza molto basse, il testo della Costituzione e gli atti comunitari si fermano a sequenze di 6 o 7 complementi.

4.2.3.6 La subordinazione

Il livello di annotazione sintattica a dipendenze si è inoltre dimostrato affidabile per rintracciate nei corpora giuridici esaminati i modi e le forme della subordinazione. L'analisi si è in particolare concentrata sul monitoraggio dei seguenti tratti:

- la distribuzione media delle **frasi** per periodo, calcolata sulla base della distribuzione delle teste verbali per periodo;
- la proporzione di principali e subordinate nel corpus;
- i tipi di subordinate (implicite e esplicite) nel corpus.

La media di frasi per periodo

L'analisi della distribuzione del numero medio di frasi presenti in un periodo è uno dei tratti qui monitorati dal momento che suggerisce alcune preliminari informazioni sul rapporto tra paratassi e ipotassi. Il dato è stato ricostruito calcolando le occorrenze di teste verbali in un periodo.

In questo caso, l'obiettivo era quello di cercare una conferma quantitativa di quanto osservato da Garavelli (2001, p. 100) riguardo al fatto che "una statistica delle strutture sintattiche impiegate rivela abbastanza alto il numero delle presenze di enunciati monoproposizionali".

I risultati riportati nella Tabella 4.15 ne sono infatti una parziale conferma. AMB mostra di contenere un numero medio di frasi per periodo inferiore a quello di Rep, ma uguale a quello di 2Par. Inoltre, gli atti amministrativi risultano essere la tipologia di testo giuridico con una distribuzione superiore alla media dell'intero corpus.

Questi dati sono di maggiore interesse, tuttavia, se intrecciati con quelli ricavati dal monitoraggio di una serie di altri tratti presi in considerazione in questo studio: quelli relativi cioè alla tipologia di frasi contenute nei periodi dei corpora analizzati, con un particolare riguardo *i)* alla proporzione di frasi principali e subordinate e *ii)* al tipo di subordinate presenti. Sono questi infatti i tratti rintracciati e discussi nei successivi paragrafi.

	Media di frasi
AMB	1,95
Rep	2,34
2Par	1,95
AMBamm(Stato)	**2,59**
AMBamm(Regione)	2,23
AMBamm(Europa)	2,18
AMBnorm(Europa)	2,14
AMBnorm(Stato)	1,81
COST	1,47
AMBnorm(Regione)	1,25

Tabella 4.15: Confronto dei valori relativi alla media di frasi per periodo nei testi normativo–amministrativi e giornalistici.

La proporzione di principali e subordinate

Il monitoraggio di questo parametro è stato condotto tenendo in considerazione il rapporto, all'interno di ciascuno dei corpora in esame, tra *i)* le radici verbali (corrispondenti alle frasi pricipali) e *ii)* le frasi sottocategorizzate o quelle con valore di modificazione temporale, causale, locativo, ecc... dipendenti da una testa verbale (corrispondenti alle frasi subordinate).

Ad esempio, nel seguente periodo, la cui struttura sintattica è graficamente rappresentata nella Figura 4.11, la radice (root) verbale dell'intero periodo (*autorizzato*) è stata considerata una frase principale; la frase *ad apportare, con propri decreti, le variazioni di bilancio occorrenti per l'attuazione dei commi 17 e 18*, sottocategorizzata dal verbo reggente (legata cioè da una relazione di dipendenza 'arg'[19] al verbo *autorizzato*), è stata considerata una frase subordinata:

- *Il Ministro dell'economia e delle finanze è autorizzato ad apportare, con propri decreti, le variazioni di bilancio occorrenti per l'attuazione dei commi 17 e 18.*

Oppure, è anche il caso del seguente periodo (la cui struttura sintattica a dipendenze è riportata nella Figura 4.12), dove la radice verbale *tenuto* è

[19]Per chiarezza si ricorda che, sulla base dello schema di annotazione sintattica a dipendenze adottato, la relazione 'argument' ('arg') è la relazione tra una testa verbale o nominale e una frase completiva non soggetto (sia essa infinitiva o meno).

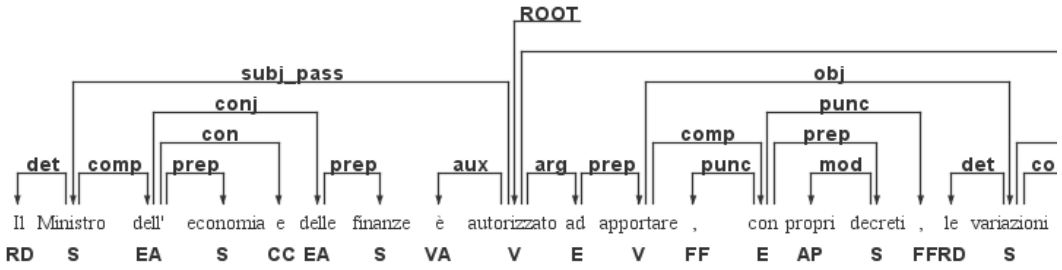


Figura 4.11: Un esempio di frase principale e di subordinata implicita sottocategorizzata dal verbo reggente *autorizzato*.

stata calcolata come frase principale e la frase *se la Commissione accetta il ritiro della notifica*, legata da una relazione di dipendenza 'mod'[20] alla radice verbale del periodo (*tenuto*), è annoverata tra le frasi subordinate:

- *Se la Commissione accetta il ritiro della notifica, il notificante non è più tenuto a rispettare i requisiti di cui al paragrafo 3.*

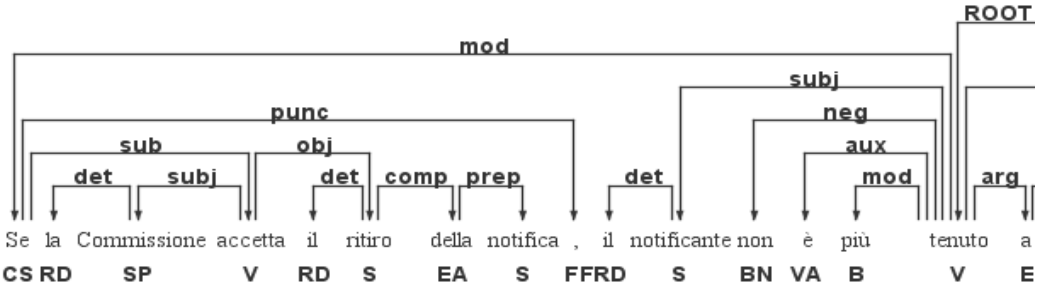


Figura 4.12: Un esempio di frase principale e di subordinata esplicita dipendente dal verbo reggente *tenuto*.

I risultati del monitoraggio di questo tratto sono riportati nella 4.13, che mostra il diverso rapporto tra la percentuale di frasi principali e subordinate

[20]Si ricorda che la relazione 'modifier' ('mod') è la relazione tra una testa e il suo modificatore; tale relazione copre modificatori di tipo frasale, aggettivale avverbiale e nominale.

nei vari corpora analizzati. L'intero corpus giuridico contiene una percentuale nettamente inferiore di subordinate (pari al 25,58% di tutte le frasi presenti) rispetto a Rep (33,93%) e inferiore a 2par (27,33%), sebbene con una differenza meno marcata.

Tra tutti i corpora giuridici, le collezioni di atti normativi regionali e statali, insieme alla Costituzione, dimostrano di essere quelli con la percentuale minore di frasi subordinate.

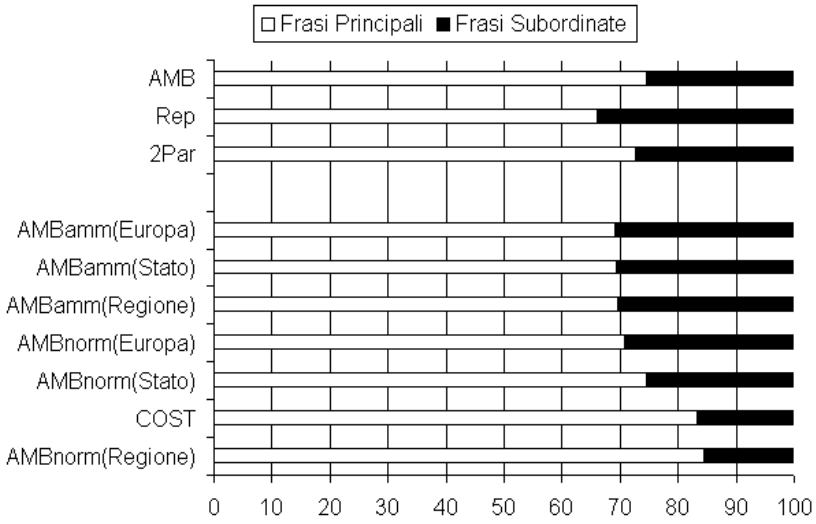


Figura 4.13: La distribuzione percentuale di frasi principali e subordinanti nei testi normativo–amministrativi e giornalistici.

Subordinate implicite ed esplicite

Lo stato della subordinazione nei testi giuridici è stato ulteriormente studiato monitorando la distribuzione dei diversi tipi di subordinate implicite ed esplicite.

L'obiettivo era quello di trovare una conferma in quanto osservato da Garavelli (2001, pp. 161-162) a proposito dei più vistosi "stereotipi sintattici" dei testi giuridici, tra i quali vengono annoverati le "sovraestensioni dell'infinito in frase completiva" e in generale l'uso di subordinate con l'infinito, laddove sarebbe possibile scegliere tra una subordinata di forma implicita e una di forma esplicita con il verbo di modo finito. In entrambi i casi viene

fatto notare da Mortara Garavelli che "il loro uso risponde alla tendenza alla riduzione sintattica; che è un tendere alla sintesi [...] sintesi strutturale, che non vuol dire eliminazione della prolissità su altri piani dell'espressione, e sul livello dell'organizzazione (o forma) del contenuto".

Sulla base dello schema di annotazione sintattica a dipendenze di partenza, l'analisi è stata condotta considerando, tra le frasi subordinate calcolate come spiegato nel paragrafo precedente, subordinate **implicite** le frasi dipendenti da una testa verbale (sottocategorizzate o con valore di modificazione temporale, causale, locativo, ecc...) e introdotte da una preposizione. Ne è un esempio il periodo riportato nella Figura 4.11 (e discusso nel precedente paragrafo), dove la frase ***ad*** *apportare, con propri decreti, le variazioni di bilancio occorrenti per l'attuazione dei commi 17 e 18*, legata da una relazione di dipendenza 'arg' alla radice verbale (*autorizzato*) del periodo, è introdotta da una preposizione (*ad*).

Sono state invece considerate subordinate **esplicite** le frasi dipendenti da una radice verbale e introdotte da una congiunzione subordinante. Ne è un esempio il periodo riportato nella Figura 4.12 (e discusso nel precedente paragrafo), dove la frase ***se*** *la Commissione accetta il ritiro della notifica*, legata da una relazione di dipendenza 'mod' alla radice verbale (*tenuto*) del periodo è introdotta da una congiunzione di tipo subordinante.

I risultati di questa analisi riportati nella Figura 4.14 dimostrano come tra i due tipi di subordinate quelle implicite siano maggioritarie nell'intero corpus di testi giuridici, dove costituiscono il 72,79% di tutte le subordinate calcolate, rispetto a Rep (67,87%). Rispetto a 2Par, dove le subordinate implicite sono il 74,68%, AMB mostra una distribuzione percentualmente inferiore di 1,89 punti.

È interessante qui suggerire come questi dati possano essere messi in relazione con la diversa distribuzione di preposizioni e di congiunzioni di tipo subordinante presa in esame nei Paragrafi 4.2.2.3 e 4.2.2.4. In quell'occasione, sulla base del livello di annotazione morfosintattica del testo, era stato fatto osservare come l'intero corpus giuridico si caratterizzassee per una maggiore distribuzione percentuale di preposizioni e una minore di congiunzioni subordinanti rispetto ai corpora giornalistici.

Un tale dato si riflette dunque nella diversa distribuzione dei tipi di subordinate: in AMB sono maggioritarie quelle implicite, introdotte da una preposizione, e sono minoritarie quelle esplicite, introdotte da una congiunzione di tipo subordinante.

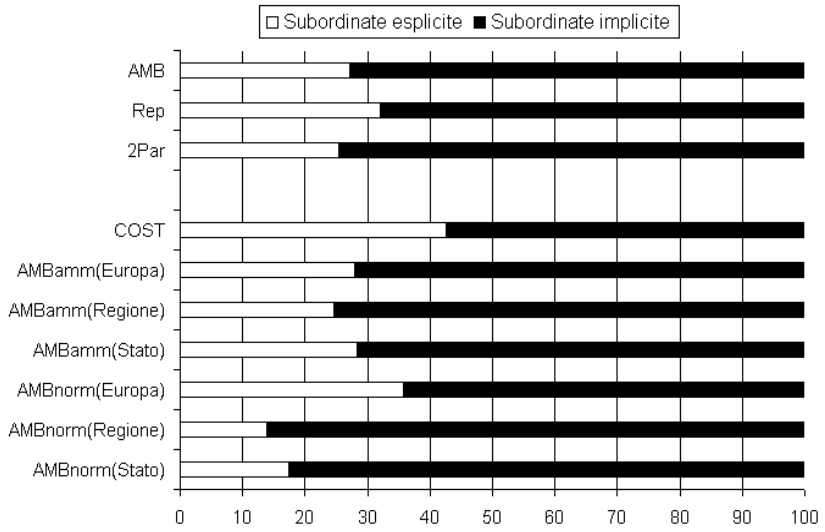


Figura 4.14: La distribuzione percentuale di subordinante esplicite ed implicite nei testi normativo–amministrativi e giornalistici.

Riflessioni analoghe si possono condurre anche rispetto alla distribuzione dei due tipi di subordinate nei vari corpora giuridici analizzati. I testi regionali e statali normativi (soprattutto) e amministrativi, che a livello morfosintattico avevano dimostrato di possedere una maggiore percentuale di preposizioni e una minore di congiunzioni subordinanti rispetto agli altri testi giuridici, sono anche quelli con una percentuale maggiore di frasi subordinate implicite e una minore di subordinate esplicite. Al contrario, la Costituzione, caratterizzata da un profilo morfosintattico opposto rispetto a questa tipologia di testi giuridici, a livello sintattico dimostra di essere il testo la percentuale maggiore di subordinate esplicite e la minore di subordinate implicite.

Si può dunque concludere che, sebbene i risultati del monitoraggio di questo tratto andrebbero ulteriormente studiati soprattutto in rapporto alle diverse scelte lessicali (a livello verbale) operate nelle varie tipologie di testo, tuttavia i dati sin qui ottenuti sono una conferma quantitativa di quanto osservato da Bice Mortara Garavelli a proposito della maggiore tendenza alla subordinazione implicita caratteristica dei testi giuridici.

4.2.4 Le caratteristiche lessicali

Sulla base del livello di lemmatizzazione e di annotazione morfosintattica è stato possibile monitorare i corpora rispetto alle loro caratteristiche lessicali. Il profilo lessicale dei testi in esame è stato in particolare ricostruito grazie al monitoraggio dei tratti discussi in quanto segue.

4.2.4.1 La densità lessicale

Ottenuti sulla base dell'annotazione morfosintattica automatica, i dati relativi al monitoraggio di questo tratto sono finalizzati a mettere in luce il rapporto tra lessico referenziale (le parole 'piene') e lessico funzionale (le parole 'vuote'). Come ricordato da Simone (1996), valori elevati di densità lessicale sono rintracciabili in testi altamente informativi, dal momento che parole 'piene' portano più informazione di quelle funzionali 'vuote'[21].

I valori sono qui stati calcolati come il rapporto tra la proporzione di nomi, verbi, avverbi e aggettivi (le parole 'piene') presenti nel corpus il totale dei tokens.

	Densità lessicale
AMB	0,545
Rep	0,567
2Par	0,565
AMBamm(Europa)	**0,560**
COST	0,556
AMBamm(Regione)	0,547
AMBamm(Stato)	0,543
AMBnorm(Europa)	0,543
AMBnorm(Regione)	0,539
AMBnorm(Stato)	0,527

Tabella 4.16: Confronto dei valori di densità lessicale dei testi normativo–amministrativi e giornalistici.

I risultati riportati nella Tabella 4.16 mostrano come l'intero corpus giuridico abbia valori di densità lessicale leggermente inferiori sia a Rep sia a

[21] Ad esempio, il monitoraggio del LIP (Lessico di frequenza dell'Italiano Parlato) (De Mauro, 1993) ha dimostrato che la lingua parlata si caratterizza per una maggiore povertà lessicale rispetto alla lingua scritta.

2Par. L'intuizione è che il dato sia in parte legato alla preponderante presenza nei testi giuridici di preposizioni (parole 'vuote'), come messo in luce nel Paragrafo 4.2.2.

Tuttavia la differenza dei valori (pari a 0,02 punti percentuali) tra le due tipologie di testi non si può considerare statisticamente significativa. Il suggerimento è che tali dati andrebbero intrecciati con quelli relativi al tipo di lessico usato nei diversi corpora.

4.2.4.2 La ricchezza lessicale

A partire dal livello di lemmatizzazione automatica del testo è stato possibile misurarne la ricchezza lessicale. L'interesse di monitorare i corpora giuridici rispetto a questo tratto è legato al fatto che esso rappresenta uno degli aspetti rispetto ai quali la "Guida per la redazione degli atti amministrativi", usata come riferimento in questo studio[22], fornisce alcuni suggerimenti. A proposito dell'uso ricco e variegato di terminologia, si raccomanda infatti di "usare sempre il medesimo termine per esprimere uno stesso concetto; alternare termini diversi per indicare lo stesso concetto al fine di evitare le ripetizioni pu generare confusione e ambiguità".

Il grado di ricchezza lessicale è stato qui calcolato sulla base del rapporto tipo/unità (Type/Token Ratio, d'ora in avanti TTR). Misura ampiamente utilizzata in statistica lessicale, la TTR consiste nel calcolare il rapporto tra il numero di parole tipo in un testo, il 'vocabolario' di un testo (V_c), e il numero delle occorrenza delle unità del vocabolario nel testo (C):

$$0 \leq |\frac{V_c}{C}| \leq 1$$

I valori di TTR oscillando tra 0 e 1 indicano se il vocabolario di un testo è poco vario (valori vicini a 0) o molto vario (valori vicini a 1).

Essendo un indice sensibile alla lunghezza del testo, è stato calcolato nei diversi corpora su porzioni di testo della stessa lunghezza. I risultati di TTR riportati su due diverse porzioni di testo, pari rispettivamente a 1000 e 15000 tokens, dimostrano infatti come i valori cambino al variare della lunghezza del testo, ovviamente diminuendo al crescere della porzione di testo considerato.

Come mostra la Figura 4.15, nella prima porzione di 1000 tokens l'intero corpus giuridico ha valori di TTR più bassi (pari a 0,35) sia di Rep (0,41) sia di 2Par (0,37), sebbene in questo caso la differenza non sia statisticamente

[22]Vedi Paragrafo 4.1.1.

significativa. La stessa tendenza si mantiene anche per la seconda porzione di 15000 tokens della quale è stata calcolata la TTR. Inoltre, sebbene non ci siano marcate differenze tra le varie tipologie di testi giuridici esaminati, gli atti amministrativi statali sono i testi meno lessicalmente ricchi.

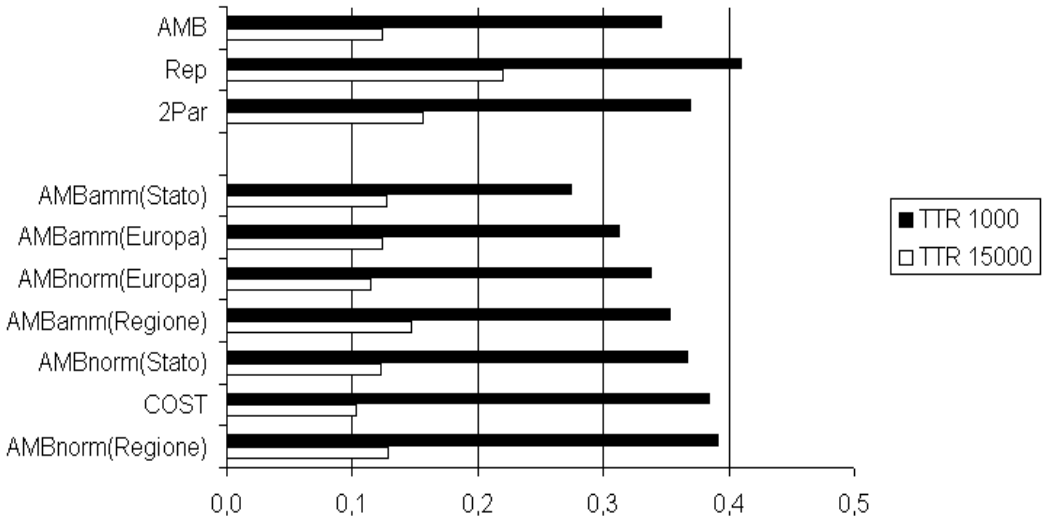


Figura 4.15: Confronto dei diversi valori di TTR in porzioni di 1000 e 15000 tokens nei testi normativo–amministrativi e giornalistici.

Questi dati andrebbero ulteriormente studiati soprattutto in relazione alla doppia tipologia di termini tipicamente presente nei testi giuridici, espressione delle due principali componenti semantiche in essi contenute: quella relativa alla realtà giuridica e quella relativa alla realtà extragiuridica a cui si fa riferimento nei testi[23].

È questo il motivo per cui i risultati ottenuti dal monitoraggio di questo tratto sono da considerarsi al momento preliminari e necessitano di essere approfonditi in futuro. Ciononostante, è interessante far notare come essi siano in linea con quanto osservato nello studio di Nystedt (1999), dove un confronto tra il lessico contenuto in una serie di testi giuridici (direttive comunitarie di diversi domini, la Costituzione italiana, il Codice Civile del 1942 e lo Statuto della regione Abruzzo) e in giornali e romanzi ha permesso di mettere in luce che *i)* i testi giuridici sono in generale più lessicalmente

[23]Sulla base della distinzione di Belvedere (1994a) esposta nel Paragrafo 2.2.1.

poveri di quelli rappresentativi della lingua comune e che *ii)* la Costituzione si contraddistingue per una maggiore ricchezza lessicale rispetto ad altre tipologie di testi giuridici.

4.2.4.3 La distribuzione del lessico rispetto al Vocabolario di Base

Anche l'interesse per il monitoraggio dei corpora rispetto a questo tratto è legato ad alcuni suggerimenti in proposito forniti dalla "Guida per la redazione degli atti amministrativi". In essa si raccomanda infatti di "scegliere le parole del vocabolario di base, preferendole a quelle più rare, dato che sono più diffuse e dunque note a tutti i parlanti".

Inoltre, l'obiettivo in questo caso era quello di trovare una conferma di quanto osservato da De Mauro (2006) riguardo allo "straordinario impegno dei *Costituenti*", al loro "non comune impegno linguistico", nell'uso di un'elevata percentuale di lessico appartenente al Vocabolario di Base nella redazione della Costituzione italiana.

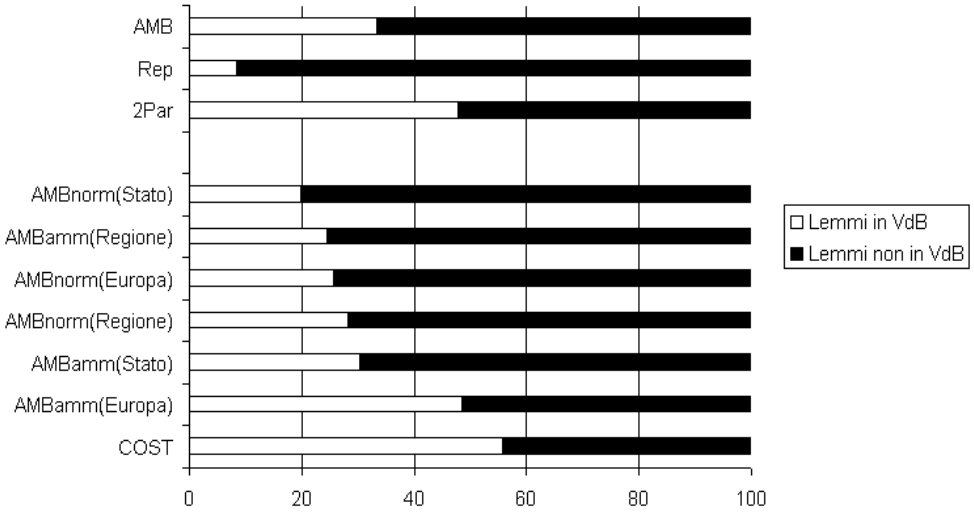


Figura 4.16: Confronto della diversa percentuale di appartenenza al Vocabolario di Base (VdB) dei lemmi contenuti nei testi normativo–amministrativi e giornalistici.

Sono questi dunque i motivi per cui si è qui scelto di monitorare i corpora considerati rispetto *i)* alla percentuale di parole appartenti al Vocabolario

di Base (d'ora in avanti VdB) del "Grande dizionario italiano dell'uso" e *ii)* al modo in cui essi si distribuiscono rispetto ai tre repertori d'uso: Lessico Fondamentale, Lessico ad Alto Uso e Lessico ad Alta Disponibilità[24].

I risultati del monitoraggio del primo tratto sono riportati nella Figura 4.16 che mostra come l'intero corpus giuridico abbia una percentuale di lemmi parte del VdB (pari al 33,28% di tutti i lemmi di AMB) nettamente superiore a Rep (8,50%), ma inferiore tuttavia a 2Par (47,78%). In particolare, tra tutti i testi giuridici, la Costituzione italiana è il testo caratterizzato dalla percentuale maggiore di lemmi parte del VdB (pari al 55,66%), con valori che superano addirittura quelli di 2Par.

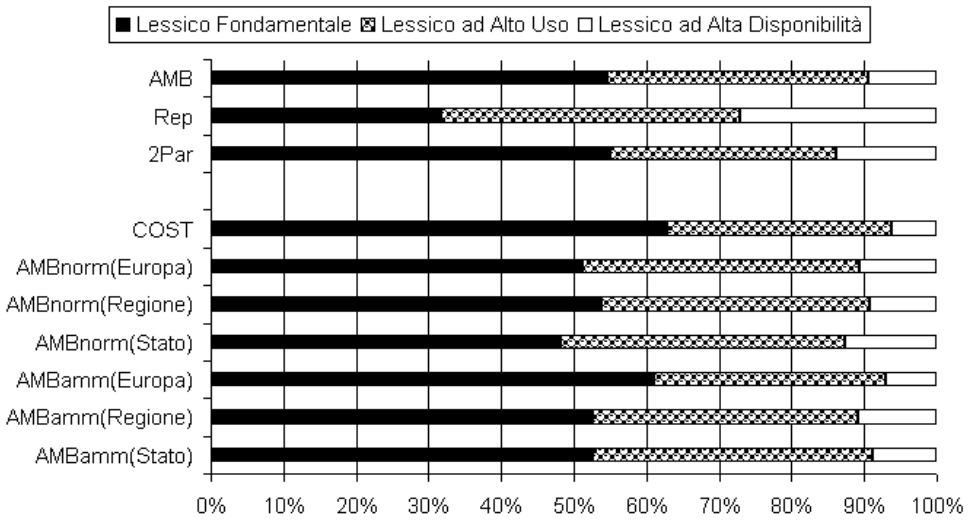


Figura 4.17: Confronto della diversa distribuzione rispetto ai repertori d'so del VdB nei testi normativo–amministrativi e giornalistici.

I risultati del monitoraggio del secondo tratto qui considerato, relativo alla distribuzione dei lemmi appartenenti al VdB nei tre repertori d'uso, sono riportati nella Figura 4.17. In essa si vede chiaramente come l'intero corpus giuridico abbia una percentuale di lemmi appartenenti al Lessico Fondamentale superiore a quella di Rep e pressoché uguale a quella di 2Par.

[24]Le parole appartenenti al VdB e ai tre repertori d'uso sono state calcolate in termini di lemmi.

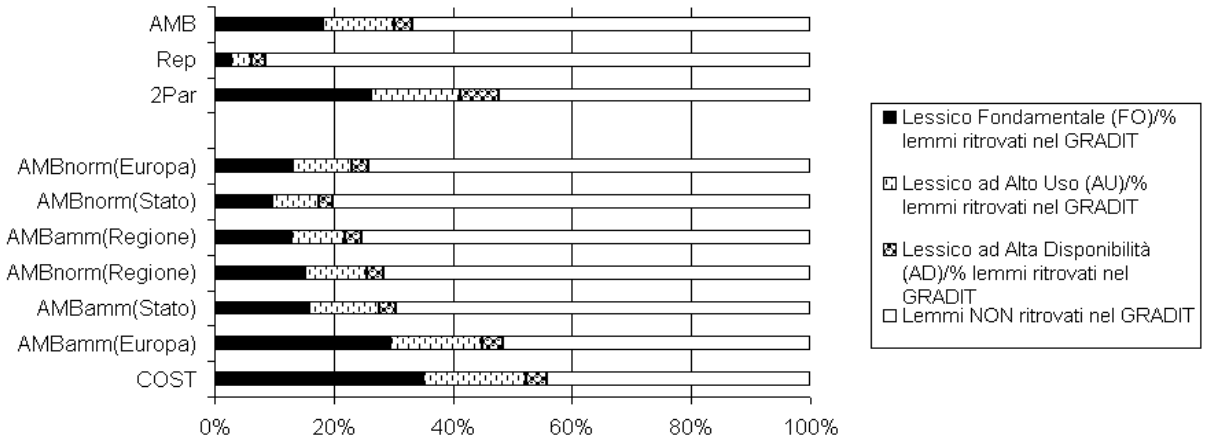


Figura 4.18: La diversa distribuzione dei lemmi appartenenti ai repertori d'uso sul totale dei lemmi del VdB nei testi normativo–amministrativi e giornalistici.

I dati sono ancora più significativi se intrecciati con quelli riportati nella Figura 4.18, dove la percentuale di appartenenza al Lessico Fondamentale, Alto Uso, Alta Disponibilità è messa in rapporto al totale di lemmi appartenenti al VdB.

I dati così raccolti ci restituiscono un corpus giuridico caratterizzato da una percentuale di Lessico Fondamentale (pari al 18,17% del totale di lemmi presenti in AMB) nettamente superiore a quella di Rep (2,70%), ma inferiore alla percentuale di 2Par (26,32%). Tali dati permettono inoltre di far notare come sia la Costituzione il testo giuridico con la presenza maggiore di lemmi appartenenti al Lessico Fondamentale, con una percentuale pari al 34,95%, superiore alla media di AMB e anche alla distribuzione di 2Par.

I risultati ottenuti dal monitoraggio di questi due tratti sono da considerarsi al momento preliminari e aprono la strada a studi futuri. Tra questi, l'indagine volta a verificare se l'elevata occorrenza nei testi giuridici di lemmi contenuti nel VdB, parte soprattutto del Lessico Fondamentale, possa essere un indizio della riconosciuta tendenza al "riuso specialistico di termini del linguaggio ordinario" (Garavelli, 2001, p. 11).

4.3 Considerazioni conclusive

Come discusso nell'introduzione a questo capitolo, la metodologia di monitoraggio linguistico qui descritta si proponeva due obiettivi principali: *i)* quello di offrire gli strumenti per condurre un'indagine quantitativa delle principali caratteristiche rintracciabili in testi giuridici volta a ricostruirne il profilo linguistico e *ii)* quello di porre le basi per il futuro sviluppo di uno strumento a supporto delle attività di verifica della redazione 'chiara, semplice e comprensibile' di un atto normativo–amministrativo e di un indicatore del livello di leggibilità di testi giuridici basato sul monitoraggio linguistico.

Nel tracciare le considerazioni conclusive di quanto esposto sin qui, sono questi infatti i due principali aspetti trattati nei due successivi paragrafi.

4.3.1 La ricostruzione del profilo linguistico dei testi giuridici

Il primo risultato sul quale si intende focalizzare l'attenzione riguarda l'affidabilità della metodologia di monitoraggio dei testi giuridici messa a punto e sperimentata in questo studio.

Come fatto notare in più punti nei paragrafi precedenti, la scelta di partire dal livello di annotazione linguistica automatica del testo per rintracciare alcuni dei più significativi tratti morfosintattici, sintattici e lessicali del corpus di atti normativo–amministrativi esaminato si è rivelata affidabile per ricostruirne un articolato profilo linguistico. La dimostrazione è tanto più evidente se si considera come, in più di un caso, una tale strategia abbia permesso di fornire una conferma quantitativa degli studi precedentemente realizzati in modo manuale.

Inoltre, l'approccio comparativo all'analisi dei testi costituisce un ulteriore aspetto centrale e innovativo di questo studio, da mettere in particolare evidenza. Esso ha permesso infatti di individuare una serie di interessanti similarità e/o differenze tra *i)* testi rappresentativi della lingua del diritto e della lingua comune e tra *ii)* le varie tipologie di testi giuridici considerati.

È stato così possibile rilevare, in primo luogo, come l'intero corpus di testi giuridici abbia comportamenti linguistici diversi da quelli rintracciati nei due corpora di testi giornalistici di riferimento, qui considerati come rappresentativi di due varietà di lingua comune: *i)* quella rappresentata dagli articoli del quotidiano "La Repubblica", leggibile e comprensibile per un ampio pubblico di lettori e *ii)* quella rappresentata dagli articoli di "Due Parole", il mensile

appositamente scritto per contenere testi leggibili e comprensibili per persone con un basso livello di alfabetizzazione e con ridotte capacità cognitive. In questo senso, dunque, il confronto tra i testi giuridici e questi due diversi corpora di prosa giornalistica è stato esplicitamente finalizzato a mettere in luce fino a che punto la lingua del diritto, che dovrebbe essere in principio comprensibile da tutti, lo è effettivamente.

Il monitoraggio comparativo ha rivelato che l'intero corpus giuridico contiene in generale caratteristiche morfosintattiche, sintattiche e lessicali più simili a quelle rintracciate negli articoli di "La Repubblica" (Rep) che alle caratteristiche dei testi di "Due Parole" (2Par). In linea con i criteri di scelta dei due corpora di riferimento, è dunque possibile affermare che i testi giuridici sono scritti in una lingua più difficile di quella appositamente pensata per essere di "facile lettura".

Riassumendo qui i tratti più significativi che sono stati monitorati, è interessante far notare che già rispetto al calcolo della lunghezza media dei periodi, i corpora giuridici hanno rivelato di contenere periodi più lunghi soprattutto rispetto a quelli contenuti di 2Par.

A partire dal livello di annotazione morfosintattica, è poi stato possibile verificare come essi si contraddistinguano per *i)* una percentuale maggiore di preposizioni e di sostantivi, *ii)* una minore di verbi, ricorrenti soprattutto nella forma participiale e *iii)* una percentuale minore di congiunzioni di tipo subordinante. In questo caso era stato, tuttavia, fatto notare come per quanto riguarda la distribuzione percentuale dei sostantivi e delle congiunzioni subordinanti i testi giuridici si avvicinassero di più alle tendenze riscontrate in 2Par che a quelle di Rep.

Ma è sulla base del livello di annotazione sintattica che sono stati raccolti i dati più significativi. È, in particolare, il monitoraggio dei corpora rispetto alla lunghezza media delle relazioni di dipendenza sintattica a rappresentare il tratto sintattico rispetto al quale il corpus di testi giuridici differisce maggiormente dai testi giornalistici. Il corpus si contraddistingue infatti per periodi caratterizzati da una distanza massima tra la testa sintattica e il suo dipendente, legati da una relazione di dipendenza, cinque volte superiore (in media) ai valori riscontrati nei testi di riferimento. Tradizionalmente considerato uno dei comportamenti sintattici maggiormente responsabili dei principali problemi di complessità (e comprensibilità) di un testo[25], una tale caratteristica è annoverata in questo studio tra le più evi-

[25]Vedi in particolare le teorie psicolinguistiche e cognitiviste basate sul calcolo dei mag-

dente peculiarità della realizzazione linguistica del contenuto informativo dei periodi giuridici[26].

Il profondo livello di incassamento gerarchico delle relazioni di dipendenza all'interno dell'albero sintattico di un periodo e, il caso particolare, delle lunghe sequenze di complementi preposizionali dipendenti da teste nominali, gerarchicamente incassati, sono altri due tra i tratti sintattici monitorati rispetto ai quali i testi giuridici hanno dimostrato di avere un comportamento nettamente diverso da quello dei testi giornalistici. Come ricordato da Dell'Orletta et al. (2011), anche in questo caso entrambi i tratti sono forti indicatori linguistici di complessità testuale. I dati ottenuti sono pertanto una conferma quantitativa di quanto affermato da Garavelli (2001) riguardo al fatto che sono i "complementi del nome a marcare sintatticamente (e testualmente) gli enunciati" giuridici, rappresentando una fonte considerevole di "complicazioni strutturali". Per questo motivo, sulla scia delle considerazioni di Bice Mortara Garavelli relative alle possibili difficoltà interpretative che derivano da sequenze concatenate di "complementi del nome", nel Paragrafo 7.4 è discusso, grazie ad alcuni esempi, come un tale comportamento sintattico possa influenzare l'organizzazione del contenuto semantico di un periodo giuridico.

È tuttavia interessante qui far notare che per quanto riguarda la subordinazione i testi giuridici hanno dimostrato di avere un comportamento più simile a quello di 2Par che a quello di Rep. Rispetto ad entrambi i corpora di testi giornalistici di riferimento, l'intero corpus giuridico contiene *i)* una media di frasi per periodo più bassa, con un numero maggiore di periodi monofrasali, *ii)* una percentuale minore di subordinate e *iii)* una percentuale maggiore di subordinate implicite (rispetto a quelle esplicite). Tuttavia, rispetto a questi tratti i testi giuridici hanno valori che si avvicinano di più

giori/minori costi di comprensione di una frase da parte dell'utente. Il dibattito, ricostruito da Fiorentino (2007) e dall'autrice ricondotto ad un contesto di semplificazione della lingua amministrativa, verte sull'idea che "il carico di informazione che la nostra memoria a breve termine è in grado di ricevere, processare e ricordare ha una misura media di 7 unità (da 5 a 9)". Pertanto "se la memoria è occupata in un compito non può svolgerne un altro in modo concomitante, deve prima liberarsi elaborando le informazioni e trasferendole alla memoria a lungo termine". Dal momento che la ricostruzione da parte dell'utente della struttura di un periodo avviene di fatto sulla base della ricostruzione delle relazioni di dipendenza, è la grande distanza tra la testa sintattica e il suo dipendente legati da una relazione di dipendenza a costituire uno dei maggiori ostacoli alla comprensione del periodo, aumentando i costi cognitivi.

[26]Vedi Paragrafo 7.4.

a quelli riscontrati in 2Par.

L'interpretazione di questi dati è complessa e necessita di futuri approfondimenti. Come ricordato da Piemontese (1996, p. 143), in un testo "è preferibile la costruzione coordinata" dal momento che "le frasi caratterizzate dalla coordinazione sono sintatticamente e semanticamente autonome, cioè costituiscono frasi grammaticalmente compiute e dotate di senso proprio". Inoltre, "sul piano della comprensibilità del testo la costruzione coordinata appare meno problematica di quella subordinata".

Tuttavia, come fatto notare da Garavelli (2003), non sempre "un discorso costruito paratatticamente [è] più semplice, e perciò più leggibile e più comprensibile di un discorso costruito ipotatticamente". Questo perché "c'è un fatto che giustifica l'impiego di strutture ipotattiche: la loro attitudine a rendere comprensibile l'ordine gerarchico dei 'pezzi' che compongono un ragionamento". Il rischio di preferire strutture coordinate, abusando nell'uso di "connettivi espliciti: ad esempio [...] *perciò, quindi, tuttavia, di conseguenza*", è quello, secondo Garavelli, di "provocare ingombri concettuali non minori di quelli che si hanno quando si esagera nelle costruzioni ipotattiche 'in verticale' ".

Inoltre, anche rispetto alle caratteristiche lessicali, i testi giuridici dimostrano di avere un comportamento più simile a quello di 2Par che a quello di Rep. In particolare, essi *i)* pur essendo meno lessicalmente ricchi dei testi di riferimento, hanno valori più vicini a quelli di 2Par e *ii)* contengono una percentuale di lemmi appartenenti al Vocabolario di Base e di questi una percentuale di lemmi del Lessico Fondamentale inferiore a quella di 2Par, ma superiore comunque a quella di Rep.

Dal monitoraggio comparativo tra le diverse tipologie di testi giuridici esaminati è stato inoltre possibile ottenere i seguenti risultati:

- il profilo linguistico degli atti amministrativi si distingue da quello degli atti normativi, caratterizzandosi per tratti morfosintattici, sintattici e lessicali particolarmente simili a quelli rintracciati in Rep e annoverati tra quelli maggiormente responsabili di un basso livello di leggibilità (e comprensibilità) di un testo. I risultati sono infatti in linea con quelli esposti da Piemontese (2000) nel suo studio finalizzato a calcolare la leggibilità di un corpus di testi normativi e amministrativi utilizzando l'indice Gulpease. Basato sul calcolo di caratteristiche generali del testo, quali la lunghezza dei periodi e delle parole, l'indice aveva rivelato che le circolari, con una lunghezza media dei periodi maggiore rispetto

alle tipologie di testi normativi esaminati, erano la tipologia di testi caratterizzati dal più basso livello di leggibilità;

- rispetto alla tipologia di ente emittente, gli atti statali e regionali hanno comportamenti linguistici simili tra loro, che li differenziano da quelli comunitari e dalla Costituzione italiana. È interessante qui far notare che tali comportamenti riguardano la distribuzione di tratti, quali *i)* periodi mediamente più lunghi, *ii)* una percentuale maggiore di preposizioni, *iii)* relazioni di dipendenza sintattica mediamente più lunghe e organizzate in incassamenti gerarchici più profondi, *iv)* complementi preposizionali dipendenti da una testa nominale organizzati in incassamenti gerarchici mediamente più profondi, ecc..., che li rendono più simili degli altri testi giuridici a Rep, suggerendo una loro possibile caratterizzazione come atti di più 'difficile lettura';

- la Costituzione italiana presenta alcune caratteristiche che la rendono differente rispetto agli altri testi contenuti nel corpus di atti giuridici monitorati. Essa dimostra in particolare di avere un profilo linguistico che si avvicina più degli altri testi a quello di 2Par. In questo senso, i risultati del monitoraggio, restituendo un testo della Costituzione caratterizzato da *i)* periodi più brevi di quelli degli altri testi giuridici, *ii)* relazioni di dipendenza sintattica addirittura più di corte di quelle presenti in 2par, *iii)* incassamenti di complementi preposizionali meno profondi, *iv)* una maggiore percentuale di lemmi appartenti al Vocabolario di Base, ecc..., sono una conferma di quanto osservato da De Mauro (2006) circa il "non comune impegno linguistico" dei padri costituenti verso la redazione di un testo leggibile e comprensibile.

4.3.2 Due scenari applicativi

Sulla scorta dei risultati ottenuti dal monitoraggio linguistico, si vuole portare l'attenzione in questo paragrafo sui due principali scenari applicativi che si aprono come possibili sviluppi futuri di questo lavoro.

Il primo è legato ad una delle tendenze e prospettive di ricerca della linguistica di testi giuridici individuate da Garavelli (2001), quella relativa cioè al "bisogno di dettare principi generali, criteri–guida, regole ben definite per la stesura di testi legislativi [...] sentito in Paesi con differenti sistemi giuridici" (Garavelli, 2001, p. 51). Come ricordato nel Paragrafo 2.3.1, in Italia

un tale "bisogno" è storicamente legato al dibattito sulla tecnica legislativa e alle prime attività in materia di legimatica.

Ponendosi in questo ambito di ricerca, la metodologia di monitoraggio linguistico descritta nei precedenti paragrafi può essere di supporto a questo tipo di attività, offrendo un mezzo utile per verificare se le indicazioni suggerite per la redazione di atti "chiari, semplici e comprensibili" nei manuali di redazione sono state effettivamente seguite dai funzionari responsabili. I vantaggi di questo supporto sono tanto più evidenti se si considera la centralità della fase di verifica durante il processo di drafting normativo (o amministrativo).

Sebbene essa ne rappresenti una parte integrante e fondamentale, tale fase è spesso quella più tralasciata, come messo recentemente in evidenza in occasione della VII giornata REI (la "Rete per l'eccellenza dell'italiano istituzionale")[27], tenutasi a Roma il 24 aprile 2009. In quell'occasione questa scarsa attenzione era stata principalmente ricondotta alla "mancanza di uno strumento informatico in grado di svolgere in automatico una buona parte della rilevazione, così da rendere più standardizzata la rilevazione stessa ed evitare gli errori derivanti da una analisi svolta 'manualmente' "[28].

La metodologia di indagine quantitativa e automatica di aspetti del profilo linguistico di atti normativi e amministrativi si propone pertanto come un possibile superamento di tale ostacolo. Il metodo è stato infatti recentemente adottato dall'"Osservatorio per il recepimento e l'attuazione della 'Guida per la redazione degli atti amministrativi. Regole e suggerimenti"' nell'ambito delle sue attività[29]. Il contributo offerto riguarda appunto la possibilità di arrivare a definire un indice di qualità redazionale di atti redatti dalla Pubblica Amministrazione sulla base dei risultati del monitoraggio di quelle caratteristiche lessicali, morfosintattiche e sintattiche che, ritenute

[27]http://ec.europa.eu/dgs/translation/rei/giornate/atti_viirei.htm

[28]L'intero intervento relativo alla questione è di Carla Paradiso, funzionaria presso il Consiglio regionale della Toscana, ed è consultabile alla pagina http://ec.europa.eu/dgs/translation/rei/documenti/giornate/viirei_paradiso.pdf

[29]Le varie attività dell'Osservatorio, insediatosi il 1 aprile 2011, sono consultabili alla pagina http://www.pacto.it/content/view/415/1/. Esse si pongono a coronamento delle ricerche che iniziate agli inizi degli anni '90 hanno visto un gran fiorire a livello nazionale di manuali e codici scritti con l'obiettivo di portare l'attenzione sulla necessità di redarre atti normativi e amministrativi 'chiari, semplici e comprensibili'. Una sintesi aggiornata di quanto fatto sin ora è stata recentemente realizzata nell'ambito delle attività didattiche e di ricerca in materia di semplificazione del linguaggio amministrativo svolte da Michele Cortelazzo e dal suo gruppo di ricerca, consultabili alla pagina http://www.maldura.unipd.it/buro/

centrali per la buona redazione di atti amministrativi, sono rintracciate in modo automatico nel testo a partire dall'output del processo di annotazione linguistica automatica.

Il secondo scenario applicativo aperto dalla metodologia di monitoraggio linguistico qui descritta riguarda la definizione di un indice di leggibilità di testi giuridici articolato su più livelli di analisi linguistica e definito sulla base di strumenti di annotazione linguistica automatica del testo.

Come suggerito nell'ultima delle regole linguistiche per la stesura degli atti amministrivi contenute nella "Guida per la redazione degli atti amministrativi", "verificare la semplicità e la comprensibilità del testo" è l'ultimo fondamentale passo del processo di drafting. A questo scopo, si raccomanda infatti "l'impiego di programmi informatici per l'edizione e l'analisi dei testi dal punto della leggibilità".

Il contributo della metodologia di monitoraggio va appunto in questa direzione, inserendosi nell'ambito del filone di ricerche avviato negli ultimi anni e attivo a livello internazionale nel quale analisi linguistiche generate da strumenti di Trattamento Automatico del Linguaggio sono usate per misurare il livello di leggibilità di varie tipologie di testi. A differenza dei metodi sino ad oggi adottati per l'analisi automatica della leggibilità, questa seconda generazione di misuratori di leggibilità non fa affidamento unicamente su caratteristiche generali e formali del testo, quali la lunghezza della frase e la lunghezza delle parole. Le misurazioni sono condotte sulla base di parametri linguistici (lessicali, morfosintattici, sintattici) monitorati in modo automatico a partire dall'output del processo di annotazione linguistica automatica del testo di cui si vuole definire il livello di leggibilità.

Per quanto riguarda la lingua italiana, il primo e al momento unico strumento sviluppato a partire da questi presupposti è rappresentato da READ–IT, descritto nei dettagli da Dell'Orletta et al. (2011). Esso, sulla base dei risultati del monitoraggio di una serie di caratteristiche linguistiche rintracciate in un corpus a partire dall'output di strumenti di annotazione linguistica automatica, permette di calcolare la leggibilità dei testi di cui il corpus è composto classificandoli come testi di 'facile' o 'difficile' lettura. La classificazione è realizzata da un classificatore statistico che associa i testi in input (linguisticamente annotati) a due 'classi' di lettura definite a priori. Si tratta di classi formate da testi tratti dal corpus "Due Parole", considerati testi di facile lettura, e dal corpus "Repubblica", considerati testi di difficile lettu-

ra[30]. L'appartenenza ad una delle due classi è stabilita sulla base del grado di similarità tra la distribuzione di alcune delle caratteristiche linguistiche monitorate. Ad esempio, testi con valori di densità lessicale, lunghezza delle relazioni di dipendenza, lunghezza di catene di complementi preposizionali modificatori di teste nominali, ecc... più vicini ai valori di monitoraggio di "Due Parole" sono classificati come testi di facile lettura rispetto a testi che mostrano valori più simili a quelli di "Repubblica".

La metodologia di analisi illustrata nei precedenti paragrafi, volta a monitorare le similarità e le differenze della distribuzione di singoli tratti linguistici nei corpora di testi giuridici considerati rispetto alle distribuzioni in questi due corpora di testi giornalistici, rappresenta dunque un primo passo verso la definizione di un loro indice di leggibilità.

La futura direzione di ricerca che qui si apre consiste infatti nello sperimentare READ–IT, specializzandolo per misurare quanto atti normativi e amministrativi siano leggibili. L'obiettivo è cioè di adattarlo tenendo in considerazione in fase di classificazione proprio quei tratti linguistici che il processo di monitoraggio descritto in questo studio ha rivelato essere particolarmente caratterizzanti questa tipologia di testi.

Ciò rappresenterebbe una novità nell'ambito delle iniziative volte a definire una metodologia di analisi della leggibilità di testi giuridici. Come già anticipato nel Paragrafo 2.3.1, sino ad oggi attività di questo tipo fanno infatti per lo più affidamento su indici in grado di computare in modo automatico caratteristiche generali e formali di un testo. Tra i contributi più significativi per la lingua italiana, l'indice maggiormente utilizzato è l'indice Gulpease[31].

È qui interessante infine far osservare come il caso italiano non sia isolato, ma sia al contrario in linea con lo stato dell'arte nell'uso degli indici di leggibilità per il dominio giuridico. Sino ad oggi, sia in Italia sia a livello internazionale i metodi adottati non seguono infatti i più recenti sviluppi che si avvalgono di strumenti di Trattamento Automatico del Linguaggio.

[30]Per le motivazioni di questa scelta vedi il Paragrafo 4.1.3.

[31]Vedi Piemontese (1996, pp. 123–193), Piemontese e Tiraboschi (1990) e Piemontese (1999, 2000, 2001).

Parte III

Dall'annotazione sintattica a quella semantica: FrameNet per il dominio giuridico

Capitolo 5

L'accesso al contenuto di testi giuridici: un processo incrementale

L'obiettivo principale dello studio presentato in questo lavoro è quello di dimostrare empiricamente come sia possibile rendere esplicito il contenuto informativo di testi giuridici a partire dall'analisi della loro struttura linguistica. A questo scopo, l'intero processo descrittivo prende le mosse da alcune considerazioni condotte negli ambiti di ricerca nei quali, sebbene da prospettive disciplinari diverse tra loro, è stato tradizionalmente dimostrato un comune interesse verso l'idea che una completa analisi di testi giuridici debba essere articolata su più livelli di indagine[1].

La prima considerazione è quella ricordata nell'introduzione di questo lavoro e riguarda la necessità per un linguista impegnato nello studio di testi giuridici di "porsi questioni linguistiche in stretta connessione con questioni giuridiche"[2]. La riflessione di Bice Mortara Garavelli mira cioè a sottolineare l'importanza di tenere separati, per "buona norma di igiene disciplinare", "l'occhiale del giurista" e "l'occhiale del linguista"[3], ma come tuttavia i due punti di vista si debbano necessariamente intrecciare allo scopo di fornire analisi esaustive.

L'importanza di trovare una connessione tra queste due prospettive di indagine del testo giuridico è riconosciuta anche dalla scuola italiana di filo-

[1] Vedi Capitolo 1.

[2] Garavelli (2001, p. 34).

[3] Questa e le due precedenti citazioni sono tratte da Garavelli (2001, p. 4).

sofia analitica del diritto. Come messo in luce nel Paragrafo 2.2, il principale aspetto d'interesse (ai fini di questo lavoro) delle riflessioni condotte in quell'ambito di ricerca riguarda l'idea per cui l'attività di interpretazione di un testo giuridico consista in un compito di semiotica linguistica, articolato su più livelli di analisi linguistica. In una tale prospettiva, il processo di semiotica giuridica viene esplicitamente fatto coincidere con un processo di semiotica linguistica e interpretare un testo giuridico (principalmente legislativo) significa, detto con le parole di Jori e Pintore (1995, p. 318), applicare "regole semiotiche giuridiche ai tre livelli: sintattico, semantico e pragmatico (anche se i giuristi che le usano di solito non le chiamano in questo modo)".

Inoltre, come discusso nel Paragrafo 2.3, la recente attenzione dimostrata dalla comunità di ricerca in AI&Law per l'uso di strumenti di Trattamento Automatico del Linguaggio ha permesso di mettere in luce i vantaggi di basare compiti di gestione del contenuto di testi giuridici sui risultati della loro annotazione articolata su più livelli di descrizione linguistica.

Sino a questo punto l'attenzione in questo studio si è focalizzata sul livello di analisi relativo all'annotazione della struttura morfosintattica e sintattica dei testi giuridici. Oggetto delle discussioni di questa terza parte dello studio è invece la metodologia di indagine adottata per accedere in modo incrementale all'informazione in essi implicitamente contenuta.

Quali siano gli eventuali aspetti problematici da tenere in considerazione in un processo di accesso al contenuto di testi giuridici, quali siano i passi fondamentali che permettono di renderlo esplicito, come essi si articolino in una successione incrementale sono dunque gli aspetti che si intende discutere in questo capitolo.

Nei paragrafi successivi è prima di tutto presentata una ben nota peculiarità del discorso giuridico, una peculiarità che ha influenzato l'intera metodologia di accesso al contenuto: lo stretto intreccio di una componente **giuridica** e di una **extragiuridica** nel contenuto degli enunciati giuridici, il loro rifarsi cioè nello stesso momento al mondo delle norme e a quello dei fatti regolati. Tema ampiamente discusso, esso è al centro del dibattito sia in materia di rappresentazione della conoscenza di dominio (Paragrafo 5.1.1) sia di definizione dei confini del lessico giuridico (Paragrafo 5.1.2).

Tenendo conto di questo aspetto, nel Paragrafo 5.2.1 è esposto il primo passo qui considerato che permette di accedere al contenuto informativo dei testi: quello che riguarda l'identificazione e l'estrazione di terminologia da essi. Dal momento che i termini, essendo gli oggetti linguistici nei quali i concetti si istanziano, rappresentano gli elementi primari della conoscenza,

l'individuazione di quali siano quelli che descrivono pienamente un dominio di conoscenza è infatti un primo passo fondamentale verso l'accesso al contenuto testuale. Per dirla con le parole di Buitelaar et al. (2005, pp. 3–12), impegnati nella discussione di come l'estrazione terminologica costituisca il primo imprescindibile gradino di un processo di costruzione stratificata di un'ontologia di dominio, "terms are linguistic realizations of domain–specific concepts and are therefore central to further, more complex tasks".

Sulla scia di questa concezione per cui l'accesso al lessico è la chiave primaria per rendere esplicito il contenuto di un testo, nel Paragrafo 5.2.1 è descritto un metodo di estrazione automatica di terminologia da testi giuridici finalizzato a discriminare i termini **fattuali** da quelli **giuridici** in essi contenuti, metodo esemplificato nel Paragrafo 5.2.2.

Il secondo passo di accesso al contenuto è in linea con l'idea per cui un'ulteriore fase del processo di interpretazione di un testo giuridico sia quella di collocare il lessico ritenuto caratterizzante il testo "nel contesto degli enunciati", per dirla con le parole di Jori e Pintore (1995, p. 212). Nel Paragrafo 5.3 è dunque presentata la metodologia (descritta poi in dettaglio nel Capitolo 6) finalizzata a ricostruire le proprietà semantico–combinatorie tra termini attraverso un processo di annotazione semantica del testo basata sulla struttura sintattica resa esplicita dalla precedente fase di annotazione linguistica.

5.1 Considerazioni preliminari: il dibattuto rapporto tra mondo delle norme e mondo dei fatti

La necessità di accedere al contenuto informativo di testi giuridici è da sempre al centro degli studi di filosofia analitica del diritto e delle attività in materia di AI&Law. Sono queste infatti le due comunità di ricerca per le quali la semantica del discorso giuridico rappresenta un campo privilegiato di indagine. Nei due diversi ambiti di studio, l'interesse a rendere esplicita l'informazione contenuta in testi giuridici si concretizza in due diversi compiti, che possono essere considerati tra loro complementari: quello di fornire un'**interpretazione del contenuto** di un atto giuridico, attraverso un'attività di semiotica giuridica, e quello di arrivare a proporre una **rappresentazione della conoscenza** del dominio giuridico, attraverso un'attività di formalizzazione del diritto.

Il medesimo obiettivo è dunque realizzato in due diversi modi. Da un lato, come ricorda Scarpelli (1969), nella sua attività principale, quella cioè di interpretazione della legge,"evitare le questioni semantiche il giurista non può". Egli deve essere in grado di rendere esplicito il significato dei principali elementi informativi contenuti in un testo al fine di interpretarne il contenuto.

Dall'altro, come precedentemente discusso nel Paragrafo 2.3, la possibilità di fornire ad un agente computazionale la conoscenza necessaria per fare dei ragionamenti e per proporre soluzioni a problemi giuridici si basa sulla capacità di rendere computabili le principali strutture concettuali giuridiche contenute in un testo attraverso la loro esplicita modellizzazione formale. In questo senso, l'accesso al contenuto testuale è centrale anche per chi è impegnato nella definizione di metodologie di annotazione semantica di testi giuridici, finalizzate a renderne esplicito il contenuto informativo.

In entrambi i casi, tuttavia, ci si trova a doversi confrontare con una ben nota peculiarità dei testi giuridici e di quelli legislativi in particolare: il fatto cioè di essere caratterizzati da un "complesso intreccio di realtà giuridiche ed extragiuridiche" che si riflette nel loro lessico, per dirla con le parole di Belvedere (1994a).

Il tema, discusso da Jori e Pintore (1995, pp. 244–245) si rifà alla distinzione esistente in un enunciato giuridico tra "la componente semantica che le norme hanno in comune con le descrizioni", quella relativa cioè alla descrizione della realtà extralinguistica, e "quella parte dell'enunciato che indica che il comportamento in questione è inteso non come una realtà sussistente [...], ma come un modello da seguire". È quest'ultima la componente del contenuto proposizionale di un enunciato normativo che "distingue le prescrizioni dalle descrizioni". Così, ad esempio, nell'enunciato *gli automobilisti sono tenuti a fermarsi al semaforo rosso* si intrecciano la componente extragiuridica, rappresentata dall'azione 'fermarsi al semaforo rosso' che descrive un possibile comportamento umano, e la componente giuridico–prescrittiva, rappresentata dal riferimento ad una norma di comportamento prevista dal Codice della Strada che prevedere l''obbligo di fermarsi davanti ad un semaforo rosso'.

Inoltre, dal momento che i termini sono la prima istanza linguistica del contenuto, un tale intreccio di realtà ha un diretto riflesso nella composizione del lessico di un testo giuridico. La mescolanza delle due componenti normativa e fattuale in un enunciato giuridico trova infatti riscontro nel fatto che in un testo giuridico i termini tecnico–giuridici, rappresentativi del mondo del diritto, sono strettamente intrecciati con quelli rappresentativi dello specifico dominio di conoscenza regolato.

Questione al centro del dibattito teorico degli studi condotti dai linguisti, da un punto di vista più pratico essa è fonte di difficoltà soprattutto (come discusso nei due successivi paragrafi) *i)* in un'ottica di rappresentazione formale della conoscenza giuridica e *ii)* per approcci basati su di un'esplicita attenzione ai termini come principale via d'accesso al contenuto di testi giuridici.

È infatti il tentativo di fronteggiare questa commistione che ha guidato la definizione in questo studio *i)* della metodologia di estrazione automatica di terminologia descritta nel Paragrafo 5.2, espressamente finalizzata a trovare un modo per discriminare le diverse tipologie di lessico presenti in un testo giuridico, e *ii)* dell'originale modalità di annotazione semantica di testi giuridici descritta nel Paragrafo 7.3.3.

5.1.1 Il "complesso intreccio di realtà giuridica ed extragiuridica"

La discussione circa il complesso intreccio di due diverse realtà all'interno del discorso giuridico riguarda il ben noto fatto che, come ricorda recentemente Biagioli (2009, p. 28), "una legge parla di processi in cui i soggetti compiono azioni su oggetti, hanno relazioni con altri soggetti, accadono eventi ed esistono o si producono stati: li descrive e li regola simultaneamente".

Pertanto, ponendosi come obiettivo quello di rappresentare in maniera formale il contenuto informativo di un testo giuridico, rendendolo esplicito, il rischio in cui si può incorrere è quello di mischiare indiscriminatamente la realtà strettamente giuridica e quella relativa al mondo fattuale regolato. Come messa in evidenza da Breuker e Hoekstra (2004), il rischio è legato ad una questione di commistione di piani di organizzazione dell'informazione.

Un modello ben formato di rappresentazione formale della conoscenza giuridica dovrebbe essere articolato su due livelli: *i)* un livello generale (definito 'core level'), nel quale sono organizzati i concetti fondamentali della teoria del diritto (espressi da termini evocativi della realtà giuridica), la cui rappresentazione non cambia al cambiare del dominio regolato, e *ii)* un livello più specifico (definito 'domain–specific level'), che offre una rappresentazione dei principali concetti che descrivono il mondo regolato (istanziati da termini evocativi della realtà extragiuridica), i quali richiedono una loro apposita descrizione formale.

Una tale organizzazione stratificata è particolarmente vantaggiosa dal momento che consente di riutilizzare il livello di rappresentazione della realtà giuridica al cambiare del mondo dei fatti oggetto di rappresentazione. È quest'ultimo infatti che necessiterà essere di volta in volta modificato allo scopo di descrivere in maniera soddisfacente le relazioni tra le relative entità rilevanti.

La questione è in questo senso di grande interesse per la comunità di ricerca in AI&Law impegnata a costruire ontologie giuridiche, intese come sistemi di organizzazione formale della conoscenza del dominio giuridico. Nel caso in cui infatti i due tipi di conoscenza (di realtà) non siano tenuti separati, il rischio è quello di sviluppare ontologie affette da quella che viene definita da Breuker e Hoekstra (2004) "epistemological promiscuity", ontologie cioè nelle quali la prospettiva **epistemologica**, relativa alla descrizione del mondo regolato (la realtà extragiuridica), è mischiata con la prospettiva **ontologica**, relativa alle primitive di conoscenza (i concetti giuridici) fondamentali per la rappresentazione della realtà giuridica.

Purtroppo, secondo quanto affermato da Breuker e Hoekstra, molte delle ontologie giuridiche in principio costruite come 'core legal ontologies', costruite cioè per contenere unicamente un livello 'core' di rappresentazione della conoscenza, "contain for almost ninety–nine percent terms that belonged to the category 'world knowledge', i.e. the world the legal domain is about". Al contrario, invece, una 'core ontology' dovrebbe contenere solo"typical legal concepts, like norm, responsibility, person (agent), action, etc.".

Suggerendo una possibile soluzione a questo stato di cose, Francesconi et al. (2010) hanno recentemente proposto un approccio alla rappresentazione formalizzata della conoscenza giuridica basato sulla distinzione tra conoscenza tecnico–giuridica e conoscenza del mondo regolato. Il modello suggerito prevede infatti due distinti livelli di organizzazione: uno, chiamato 'Domain Independent Legal Knowledge level' (DILK), nel quale sono formalmente organizzati i concetti giuridici per lo più relativi alla funzione prescrittiva delle norme (es. divieto, permesso, dovere); un secondo, chiamato 'Domain Knowledge level' (DK), nel quale sono resi espliciti i principali concetti, e le relazioni che li legano, rappresentativi di un determinato dominio di conoscenza regolato dalle norme. L'articolazione a doppio livello è esplicitamente finalizzata per essere usata nella costruzione di sistemi di rappresentazione della conoscenza (ontologie giuridiche) che siano riutilizzabili per la modellizzazione dei diversi domini specialistici legislati.

5.1.2 La mescolanza di termini "fattuali" e giuridici

Come discusso nel Capitolo 2, in particolare nei Paragrafi 2.1.1 e 2.2.1, l'analisi della natura composita del lessico giuridico è da sempre al centro degli studi sia di linguisti sia di teorici e filosofi del diritto. Riflesso delle discussioni circa le due diverse componenti (giuridica e extragiuridica) che caratterizzano il contenuto proposizionale di enunciati giuridici, essa è connessa soprattutto con la difficoltà di stabilire i confini tra lessico tecnico–giuridico, lessico comune e il lessico delle discipline specialistiche oggetto del discorso giuridico, i termini "fattuali" secondo la definizione proposta da Belvedere (1994a).

Un tale stretto intreccio tra lessico tecnico–giuridico e lessico comune è riconducibile, secondo le riflessioni dei linguisti, ad una questione di **scelta** da parte del legislatore, la scelta "di operare nell'ambito dei valori lessicali risaputi", quelli cioè del lessico comune, rinunciando così "ad una sistemazione rigida della terminologia"[4]. Da una tale scelta deriva una lingua che ben esemplifica i noti e non lievi problemi di delimitazione tra linguaggi specialistici e lingua comune rispetto ai due assi di variazione linguistica, quello "orizzontale", relativo ai "confini disciplinari" tra linguaggi specialistici, e quello "verticale", relativo alle diverse "tipologie comunicative"[5]. Il carattere "multiforme e complesso" riconosciuto da Cortelazzo (1997) alla lingua del diritto si manifesta infatti nel fatto che essa "più delle altre fa ricorso a risemantizzazioni del lessico comune, [...] diffonde nel lessico comune i propri termini, e [...] contemporaneamente è impegnata in scambi comunicativi cui partecipano anche parlanti non specialistici".

Nell'ambito degli studi di semantica giuridica condotti da teorici e filosofi del diritto la situazione è legata piuttosto alla natura stessa della lingua del diritto e in particolare all'intrecciarsi nel discorso giuridico di più **regole d'uso** relative ai diversi tipi di realtà espresse. In quest'ottica, nella lingua del diritto, ovvero una lingua finalizzata a "dar norma alla vita comune e ad attività specialistiche di ogni genere in mille diversi aspetti" (Scarpelli, 1959), sono preservati i significati di concetti che si riferiscono sia alla realtà comune sia a quella specialistica, normata dal diritto. Nello stesso tempo tuttavia tali concetti sono ridefiniti in modo tale che la lingua del diritto costituisce "la struttura intorno alla quale se ne organizzerà l'impiego" (Scarpelli, 1959).

In un'ottica più pratica, un tale stato di cose è fonte di difficoltà messe in luce da chi vede nei termini presenti nei testi giuridici la principale via

[4] È la tesi di De Mauro (1963, pp. 426–428).

[5] Sulla base delle distinzioni di Rovere (1989).

d'accesso al loro contenuto informativo.

È il caso questo del processo di interpretazione giuridica, tanto più difficile quanto più comporta l'interpretazione di norme giuridiche che contengono termini assunti da un linguaggio specialistico; in questo caso l'aspetto problematico è connesso con una questione di attribuzione del ruolo di interprete del diritto, un ruolo che deve essere svolto "da un esperto della disciplina" o "dall'esperto di dominio", come si interrogano Jori e Pintore (1995)?

È il caso anche dei problemi connessi (come discusso nel paragrafo precedente) con la rappresentazione formale della conoscenza giuridica, delle difficoltà cioè di stabilire prima di tutto di quale realtà (giuridica o extragiuridica) siano espressione i termini che si intende considerare come spie di concetti e, una volta stabilito ciò, in quale livello di rappresentazione essi debbano essere organizzati.

Ma è anche il caso, come discusso nel paragrafo successivo, delle difficoltà che deve fronteggiare chi mette a punto approcci all'identificazione e estrazione automatica di terminologia rilevante da corpora di testi giuridici.

Ed è infine fondamentale qui ricordare come la questione sia di primaria importanza anche per la scelta di quale componente del contenuto informativo di un testo giuridico sia di maggiore interesse rendere esplicito in un processo di annotazione semantica. È infatti questa la questione affrontata nel Capitolo 7.3.3, per risolvere la quale è stata messa a punto in questo lavoro una specifica modalità di annotazione. Essa è espressamente finalizzata a distinguere la componente fattuale da quella giuridico–deontica entrambi parte del contenuto proposizionale di enunciati giuridici.

5.2 L'accesso al lessico dei testi giuridici: l'estrazione automatica di terminologia

Come anticipato nell'introduzione di questo capitolo, il primo passo per rendere esplicito il contenuto di testi giuridici è rappresentato dall'accesso al loro lessico, dall'identificazione ed estrazione cioè di quei termini in grado di costituire le spie lessicali dei concetti in essi contenuti. In una prospettiva di utilizzo di metodi di elaborazione automatica del testo, la discussa commistione di termini fattuali e giuridici è tuttavia all'origine di ben note difficoltà di identificazione e estrazione automatica di terminologia rilevante da corpora giuridici.

Come dimostrato da Lame (2005), infatti, le misure statistiche comunemente utilizzate per determinare la probabilità che un'unità lessicale presente in un corpus sia un termine rilevante per il dominio in esame non riescono ad affrontare con successo un caso particolare come quello rappresentato dai testi giuridici nei quali lessico tecnico–giuridico e lessico fattuale–specialistico sono strettamente intrecciati tra loro e con il lessico comune.

Ciò è legato al modo in cui sono progettati i sistemi di estrazione terminologica, finalizzati a individuare terminologia specialistica a partire da corpora di dominio caratterizzati da un lessico espressione di un unico dominio di conoscenza e nettamente separato da quello comune. È questo il motivo per cui, sino ad oggi, i migliori risultati dei sistemi di estrazione automatica di terminologia sono ottenuti a partire da testi, ad esempio, come quelli di letteratura biomedica che rappresentano un caso esemplare di netta separazione tra le due tipologie di lessico.

In questo senso, dunque, il dominio giuridico caratterizzato da una commistione di tipologie di termini, non sempre nettamente distinguibili tra di loro, rappresenta un caso particolarmente complesso e di difficile risoluzione. Come sperimentato e discusso da Agnoloni et al. (2009) e da Lenci et al. (2009) per la lingua italiana, i metodi tradizionali di estrazione terminologica riescono per lo più ad acquisire liste di unità terminologiche mono e polirematiche nelle quali le diverse tipologie di termini sono indiscriminatamente mischiate. Nella discussione dell'esperimento condotto da Lenci et al. (2009) viene inoltre fatto notare come il glossario finale estratto contenga più termini giuridici che fattuali. Questo è dagli autori ricondotto alla bassa frequenza (e alto rango) dei termini fattuali nel corpus di testi giuridici di partenza, in accordo con la legge di Zipf.

Con l'obiettivo di suggerire una possibile soluzione al problema, in quanto segue viene descritta una metodologia di estrazione automatica di terminologia da corpora testuali che sperimentata su un corpus di testi giuridici si è rivelata affidabile per riuscire a individuare i termini rilevanti in esso contenuti, facendo distinzione tra quelli tecnico–giuridici, quelli fattuali e il lessico comune.

5.2.1 Il metodo di estrazione automatica di terminologia

L'aspetto innovativo del metodo qui descritto[6] consiste principalmente nell'approccio di tipo **contrastivo** seguito, in base al quale l'estrazione di unità terminologiche monorematiche e polirematiche è condotta a partire dal confronto della loro distribuzione nel corpus di acquisizione rispetto a un corpus di riferimento (detto anche 'corpus di contrasto'). Ciò fa sì che la lista finale di unità terminologiche estratte contenga quelle unità che sono maggiormente rilevanti nel corpus di acquisizione rispetto (ovvero 'per contrasto') al corpus di riferimento.

Sino ad oggi, pochi sistemi di estrazione terminologica automatica sono basati su questo metodo, sebbene alcune rilevanti eccezioni siano rappresentate da Penas et al. (2001), Chung e Nation (2004) e da Basili et al. (2001). Nonostante ognuno di questi lavori abbia messo a punto strategie diverse per computare la misura della diversa rilevanza di unità terminologiche all'interno dei corpora che vengono confrontati, un assunto condiviso li accomuna: l'idea cioè che sia possibile discriminare tra termini appartenenti ad un lessico specialistico e parole della lingua comune sulla base di un'analisi contrastiva della loro distribuzione in un corpus di dominio (il corpus di acquisizione) rispetto a un corpus rappresentativo della lingua comune (usato come corpus di contrasto).

Per quanto forniscano una risposta positiva al problema di discriminare 'termini' da parole comuni, tali sistemi presentano tuttavia due limiti fondamentali *i)* per quanto riguarda il modo con cui vengono acquisite le unità terminologiche polirematiche e *ii)* per il fatto di non essersi mai confrontati con le difficoltà connesse con la distinzione automatica, all'interno di un unico corpus di acquisizione, di termini appartenenti a più di un lessico settoriale, come nel caso dei testi giuridici.

Una differenza notevole con gli altri approcci contrastivi riguarda l'estrazione delle unità polirematiche. Rispetto a questa questione i metodi menzionati possono *i)* includere nel risultato finale del processo estrattivo unità polirematiche non rilevanti ma lessicamente governate da una testa lessicale che è stata identificata come specifica per il dominio; *ii)* non includere unità polirematiche rilevanti che non sono state acquisite perché la loro testa lessicale non è stata selezionata come specifica per il dominio.

[6]Parti di quanto segue sono riprese da Bonin et al. (2010a) e da Bonin et al. (2010b).

Ad esempio, dunque, nel caso dell'estrazione terminologica condotta a partire da un corpus di articoli scientifici sul cambiamento climatico, l'unità terminologica polirematica *effetto serra* è acquisita solo sulla base della precedente identificazione dell'unità monorematica *effetto*. Di conseguenza, nel caso in cui l'unità monorematica *effetto* non sia stata selezionata come rilevante per il corpus di acquisizione, neanche *effetto serra*, l'unità polirematica di cui *effetto* è la testa, sarà estratta, sebbene significativa per il dominio. Ma se *effetto* è stato selezionato come unità monorematica rilevante, allora tutte le polirematiche con testa *effetto*, se ricorrenti nel testo, potranno essere estratte come termini di dominio a prescindere dalla loro effettiva rilevanza per il dominio.

Al contrario, il metodo qui adottato permette di considerare la rilevanza di dominio di una polirematica sulla base della sua settorialità, come elemento unico e non rispetto alla rilevanza della monorematica che ne costituisce la testa lessicale. Così, ad esempio, sono acquisite perché rilevanti per il dominio unità polirematiche come *effetto serra* o *effetto del cambiamento climatico* e saranno escluse dal risultato finale unità quali *effetto positivo* o *effetto domino*, presenti nel corpus di acquisizione ma non rilevanti per il dominio in questione.

Tale approccio trova conferma nello studio di De Mauro e Voghera (1996). Gli autori conducendo un'analisi dei lessemi complessi (LC) presenti nel "Lessico di frequenza dell'italiano parlato" (LIP), rispetto al grado di composizionalità del loro significato, a proposito dei LC appartenenti a linguaggi settoriali, concludono che "non sempre la settorialità di un LC è connessa con l'esistenza di accezioni speciali dei membri componenti, ma può derivare dal fatto che il LC assume in determinati contesti un significato globale speciale". Ciò comporta che la settorialità di un LC non è necessariamente funzione della rilevanza di dominio delle unità monorematiche di cui il LC si compone.

Questo risulta particolarmente significativo nel caso dell'estrazione di terminologia da corpora di testi giuridici caratterizzati da una lingua alquanto 'formulaica'. Le ricerche svolte da Nystedt (2000) e da Eklund-Braconi (2000) su corpora di documenti normativi europei offrono una dimostrazione empirica di ciò. In particolare, lo studio realizzato da Eklund-Braconi (2000, p. 89 e segg.) dimostra come "l'analisi della singola parola non sia sufficiente a fornire il quadro semantico completo e reale" del corpus di normativa europea in materia ambientale esaminato. Al contrario, risultati più significativi per il dominio si ottengono dall'esame di quelle parole che "sono spesso legate

tra loro in formule più o meno fisse" così da costituire "unità semantiche complete" dotate di un "significato finito e specialistico".

5.2.1.1 Le fasi del processo di estrazione

Il punto di partenza della metodologia di estrazione automatica di terminologia adottata in questo studio è la fase di annotazione morfosintattica automatica. Nella prima fase del processo di estrazione, infatti, il corpus di acquisizione viene linguisticamente annotato dal modulo di annotazione morfosintattica descritto da Dell'Orletta (2009)[7]. Dal testo così annotato, attraverso l'uso di filtri linguistici e statistici, vengono estratte due liste di potenziali unità terminologiche monorematiche e polirematiche candidate all'estrazione.

I filtri linguistici consentono di individuare all'interno del corpus di acquisizione: *i)* le potenziali unità monorematiche, sulla base della categoria morfosintattica assegnata[8]; *ii)* le potenziali unità polirematiche, sulla base di una serie di sequenze di categorie morfosintattiche rappresentative di diversi tipi di modificazione nominale. Ad esempio, da una sequenza come 'sostantivo+aggettivo' sono individuate polirematiche quali *arte contemporanea*, *moto ondoso*; da una sequenza 'sostantivo+preposizione+sostantivo' sono individuati potenziali termini quali *massa d'aria*, *licenza d'importazione*; per arrivare a sequenze complesse come 'sostantivo+preposizione+sostantivo+ preposizione+sostantivo' sulla base della quale è individuato un termine come *nuclei di condensazione di nubi*.

È qui importante sottolineare che in questa fase è possibile personalizzare il processo di estrazione rispetto *i)* alla sequenza di categorie morfosintattiche che costituiscono le unità polirematiche candidate all'estrazione e *ii)* alla lunghezza delle potenziali unità polirematiche[9]. È possibile dunque decidere di variare la tipologia di modificatori nominali prevedendo anche la presenza per esempio di unità lessicali che contengono avverbi o congiunzioni; o è possibile estrarre unità lessicali non esclusivamente nominali; o ancora, è possibile imporre diverse lunghezze massime dei termini da estrarre a seconda dei requisiti del corpus di acquisizione.

[7]Vedi il Paragrafo 3.2 per maggiori dettagli sulla fase di annotazione morfosintattica.

[8]In questo caso si tratta sempre di sostantivi.

[9]In entrambi i casi, le scelte prese possono avere effetti sulla lista finale di termini estratti, influendo sulla sua precisione e copertura.

I filtri statistici consentono poi di ordinare i termini potenziali individuati sulla base della loro rilevanza all'interno del corpus di acquisizione, attribuendo loro un valore di significatività. In particolare, la significatività delle unità monorematiche viene stabilita sulla base della loro frequenza di occorrenza all'interno del corpus di acquisizione; mentre le unità polirematiche sono ordinate sulla base del C–NC Value (Frantzi e Ananiadou, 1999), una delle misure più utilizzate nei sistemi di estrazione terminologica[10]. Il risultato di questa fase è rappresentato da una lista di unità monorematiche e polirematiche, costituite sia da termini specialistici per il dominio sia da parole comuni.

In questa fase, inoltre, dalla lista di potenziali unità terminologiche sono filtrate quelle unità che contengono locuzioni preposizionali, come ad esempio *ai sensi di*. Ciò è reso possibile grazie all'applicazione di queste prime due fasi di estrazione ora descritte per estrarre dal corpus di partenza locuzioni preposizionali date dalla sequenza 'sostantivo+preposizione+sostantivo'. In questo modo, un'unità polirematica, come ad esempio *ai sensi della legge*, con poco significato e scarsa rilevanza non sarà annoverata tra le potenziali polirematiche candidate all'estrazione a causa della sovrapposizione con la locuzione *ai sensi di*.

È da notare che l'ordinamento ottenuto sulla base dei filtri statistici utilizzati nelle prime due fasi di estrazione non permette ancora di discriminare in modo preciso tra lessico specialistico e lessico comune. Ciò avviene nella successiva fase di confronto, all'interno della quale la distribuzione di una selezione di termini candidati, effettuata sulla base dei valori di significatività ad essi assegnati, viene confrontata con la distribuzione delle medesime unità in un corpus usato come riferimento. Questo passaggio permette di riorganizzare la selezione di termini candidati all'estrazione rispetto ad un valore di contrasto calcolato da una funzione statistica[11]. Ne risulta che, ai termini più significativi per il dominio di appartenenza del corpus di acqui-

[10]Per l'estrazione di unità terminologiche polirematiche in letteratura si parte dall'assunto di base che se due o più parole formano un termine è molto probabile che nell'uso reale esse tendano a ricorrere insieme in maniera statisticamente significativa. La significatività del legame sussistente tra le parole che formano il termine viene calcolata attraverso il ricorso a misure di associazione che considerano la frequenza di co–occorrenza delle parole che compongono l'unità terminologica polirematica in relazione alle occorrenze totali delle singole parole che la formano: per menzionarne alcune, 'Mutual Information' (Church e Hanks, 1990), 'Log–likelihood' (Dunning, 1993). In questo studio si è deciso di utilizzare il C–NC Value, la misura più recentemente messa a punto.

[11]Per maggiori dettagli sulla funzione statistica utilizzata vedi Bonin et al. (2010b).

sizione sarà associato un valore di contrasto maggiore, mentre a quelli meno significativi saranno attribuiti valori più bassi.

Tendendo a valorizzare dati con bassa frequenza di occorrenza, questa funzione, oltre a favorire l'identificazione di unità polirematiche tipicamente meno frequenti in un testo, offre (come sarà discusso nel paragrafo seguente) una soluzione al problema individuato da Lenci et al. (2009) relativo alla bassa frequenza (e alto rango) del lessico 'fattuale' in corpora di testi giuridici, per questo motivo poco rappresentato in glossari terminologici estratti in modo automatico.

5.2.2 Un esempio: l'estrazione di terminologia da atti normativi comunitari

Allo scopo di fornire un esempio della metodologia di estrazione automatica di terminologia descritta, sono qui di seguito riportati i risultati di un esperimento condotto a partire dal corpus AMBnorm(Europa) composto da testi normativi comunitari[12].

Sulla base dell'approccio contrastivo adottato, è stato prima di tutto scelto il corpus di contrasto da usare come riferimento durante la fase di confronto. Due sono stati i corpora scelti, dal momento che l'obiettivo era quello di sperimentare se e in che misura la metodologia fosse affidabile per discriminare all'interno del corpus di partenza *i)* termini della lingua comune, *ii)* termini giuridici e *iii)* termini fattuali. La fase contrastiva è stata dunque di fatto reiterata due volte consecutive.

La distribuzione delle unità mono e polirematiche candidate all'estrazione è stata prima confrontata con una porzione del corpus PAROLE (Marinelli et al., 2003), un corpus di italiano contemporaneo di circa 3 milioni di parole rappresentativo del lessico comune, e in un secondo momento con corpus di 72.210 parole composto da atti normativi comunitari in materia di protezione del consumatore (d'ora in avanti chiamato CONS).

La prima fase di contrasto è stata finalizzata ad acquisire il lessico rilevante in AMBnorm(Europa) rispetto ('per contrasto') a quello comune contenuto in PAROLE; la seconda fase ha permesso di discriminare il lessico ambientale (cioè i termini 'fattuali') da quello tecnico–giuridico (cioè i termini 'tecnico–giuridici') grazie al confronto con una collezione di testi giuridici

[12]Vedi Paragrafo 4.1.2 per la descrizione del corpus.

che regolano un mondo di fatti diverso dal dominio ambientale regolato dagli atti normativi contenuti nel corpus di acquisizione AMBnorm(Europa).

Si è scelto qui di discutere unicamente le liste di termini polirematici estratte, data la maggiore significatività di questi termini rispetto a quelli monorematici in un processo di accesso al contenuto di corpora di dominio[13].

La fase di identificazione dei potenziali termini polirematici candidati all'estrazione è stata personalizzata, imponendo che venissero selezionate, sulla base dei filtri linguistici, sequenze di categorie morfosintattiche lunghe fino a 6 elementi. Una tale soglia è stata definita su basi empiriche, consentendo di estrarre sequenze complesse del tipo 'sostantivo+aggettivo+aggettivo+ preposizione+aggettivo+sostantivo' sulla base delle quali sono state individuate unità polirematiche candidate all'estrazione come *inquinamento atmosferico transfrontaliero a grande distanza*, termine rilevante all'interno del corpus AMBnorm(Europa).

Le prime 600 unità della lista di potenziali termini polirematici individuati in questa prima fase sono state ordinate per valori decrescenti sulla base del C–NC Value, la funzione statistica qui usata per misurare la significatività per un'unità polirematica di essere un termine rilevante all'interno del corpus di acquisizione. Come mostra la prima parte della Tabella 5.1[14] che riporta alcuni dei risultati ottenuti in questa fase, il filtro statistico non consente ancora di distinguere le tre tipologie di termini ricercati. All'interno della lista di unità sono infatti mischiati *i)* termini appartenenti al lessico comune come *giorno successivo alla pubblicazione*, *ii)* termini tecnico–giuridici come *parlamento europeo* o *autorità competente* e *iii)* termini del dominio ambientale regolato come *sviluppo sostenibile* o *gas ad effetto serra*, e anche *iv)* errori di estrazione come *applicazione della presente*.

È infatti la successiva fase di confronto prima con PAROLE e poi con CONS che ha permesso di operare questa distinzione. A questo scopo, la distribuzione delle 600 unità, ordinate sulla base del C–NC Value, è stata confrontata con la distribuzione delle stesse in PAROLE. Come si può vedere nella seconda parte della Tabella 5.1, sebbene rimangano ancora alcuni termini non rilevanti all'interno del corpus di acquisizione, l'ordinamento ottenuto sulla base della funzione statistica di contrasto ha permesso di elimi-

[13]Vedi sulla questione Jackendoff (1997), Nakagawa e Mori (2003) e Chung e Nation (2004).

[14]In ognuna delle tre parti della tabella sono riportati nella prima colonna la posizione del termine nella lista estratta, nella seconda il termine estratto e nella terza la tipologia di termine (comune, tecnico–giuridico, ambientale).

nare dalla lista di termini poliremati gran parte di quelli non appartenenti al lessico tecnico–giuridico o ambientale. È il caso, ad esempio, del termine *giorno successivo alla pubblicazione* che sulla base dell'ordinamento del C–NC Value occupa la 39esima posizione nella lista di termini estratti, mentre sulla base del contrasto con PAROLE è il 228esimo termine estratto.

Tuttavia, all'interno di tale lista, termini tecnico–giuridici e ambientali sono ancora mischiati. Essi sono infatti distinti grazie alla seconda fase di contrasto, condotta confrontando la distribuzione delle prime 300 unità con la loro distribuzione in CONS. I risultati, riportati nella terza parte della Tabella 5.1, mostrano come la funzione di contrasto sia riuscita a discriminare i termini del mondo ambientale (come *valore limite*, *sostanza pericolosa*, ecc...) da quelli del mondo giuridico (come *funzionamento del mercato interno*, *disposizione nazionale*, ecc...). I primi infatti occorrendo meno frequentemente (o non occorrendo affatto) in CONS hanno un valore di contrasto maggiore e sono contenuti nella prima parte della lista finale estratta da AMBnorm(Europa); i secondi, essendo condivisi dai due corpora di atti normativi comunitari, occorrono frequentemente in entrambi i corpora e hanno di consequenza valori di contrasto inferiore, posizionandosi così nell'ultima parte della lista.

Un esempio significativo è rappresentato dai termini *effetto serra* e *ravvicinamento delle disposizioni legislative*: nel primo caso, trattandosi di un termine ambientale, esso dalla 26esima posizione occupata nella lista estratta sulla base del C–NC Value passa ad essere il 37esimo termine sulla base del contrasto con PAROLE e ad essere l'ottavo termine più significativo in AMBnorm(Europa) sulla base del contrasto con CONS. Nel secondo caso, trattandosi di un termine tecnico–giuridico la situazione è capovolta: il termine passa dall'essere il 41esimo sulla base dell'ordinamento per valori di C–NC Value, all'essere il 40esimo sulla base del contrasto con PAROLE e arriva ad essere il 296esimo termine sulla base del contrasto con CONS.

Infine, un caso emblematico è quello del termine *parlamento europeo*, che occorrendo con frequenza pari in AMBnorm(Europa) e in CONS non viene estratto nella lista finale dopo il contrasto con CONS; al contrario esso è il primo termine sulla base dell'ordinamento per C–NC Value e dei valori di contrasto con PAROLE.

La valutazione quantitativa dei risultati ottenuti ha permesso di stabilire quanto un tale metodo di estrazione automatica di terminologia fosse affidabile. Essa è stata condotta confrontando la lista di unità terminologiche estratte dopo le due fasi di contrasto con *i)* il "Dizionario Giuridico" (Edizio-

ni Simone)[15], per la valutazione dei termini tecnico–giuridici e *ii)* il thesaurus "EARTh" ("Environmental Applications Reference Thesaurus")[16]. Questo ha permesso di verificare la copertura della lista estratta in modo automatico rispetto ai termini appartenenti al dominio ambientale. È poi seguita una fase di verifica manuale da parte di esperti del dominio giuridico e ambientale, i quali hanno raffinato la precedente fase di valutazione[17].

I risultati della valutazione hanno permesso di stabilire che dopo la fase di estrazione condotta sulla base del C–NC Value il 65,34% dei termini contenuti nella lista estratta è costituito da termini rilevanti in AMBnorm(Europa), di cui un 38,67% di lessico ambientale e un 26,67% di lessico tecnico–giuridico. Al termine poi della doppia analisi contrastiva le unità terminologiche ambientali aumentano fino al 43,33% e quelle del lessico tecnico–giuridico fino al 29,33%. Il che dimostra come la fase di contrasto permetta di acquisire una lista finale contenente il 72,66% di termini poliremiatici significativi per il corpus di acquisizione, con un incremento complessivo del 7,32% rispetto alla lista di termini estratti unicamente sulla base della funzione statistica.

Inoltre, l'iterazione della fase di contrasto si è dimostrata affidabile per discriminare all'interno della lista di unità terminologiche estratta quelle appartenenti alla realtà giuridica e quelle rivelatrici della realtà extragiuridica contenute in AMBnorm(Europa). Questo è chiaramente visibile nella Figura 5.1, che mostra la distribuzione dei termini del lessico ambientale e del lessico tecnico–giuridico nella lista finale di 300 unità poliremiatiche estratte (suddivisa in gruppi di 30 termini). Come si può vedere, mentre nella prima parte della lista i termini ambientali sono in maggioranza rispetto a quelli appartenenti al lessico del diritto, nell'ultima parte la tendenza si inverte.

È infine interessante far notare che l'analisi dei risultati condotta dall'esperto giuridico ha permesso di mettere in luce come il metodo di estrazione automatica permetta di acquisire in particolare alcuni dei termini che, appartenenti al dominio ambientale, sono ridefiniti negli atti normativi di AMBnorm(Europa). A questi termini infatti viene associato un valore di contrasto molto elevato e per questo sono compresi nel primo dei dieci grup-

[15]http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A

[16]http://uta.iia.cnr.it/earth.htm#EARTh%202002

[17]La valutazione dei termini del lessico tecnico–giuridico è stata condotta dalla dottoressa Angela D'Angelo della Scuola Superiore Sant'Anna di Pisa; quella dei termini del lessico ambientale dal dottor Paolo Plini dell'Istituto di Inquinamento Atmosferico, Unità di Terminologia Ambientale del CNR di Roma.

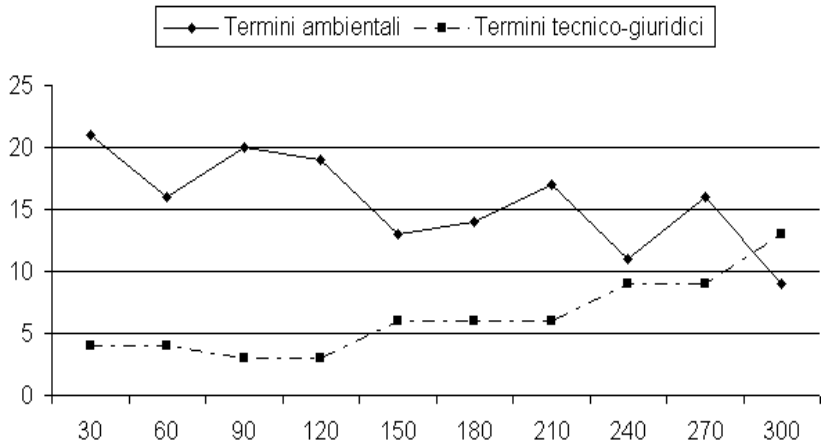


Figura 5.1: Andamento della distribuzione di termini ambientali e tecnico–giuridici all'interno della lista finale (suddivisa in gruppi di 30 termini) di termini estratti dopo le due fasi di contrasto.

pi di 30 termini nei quali (ai fini della valutazione) è stata divisa la lista finale estratta.

È il caso per esempio del termine *rifiuto pericoloso*, il cui significato è espressamente ridefinito nell'articolo 2, lettera g), del Regolamento del Parlamento europeo e del Consiglio, 25 novembre 2002, n. 2150/2002[18], contenuto in AMBnorm(Europa), come: "qualsiasi rifiuto definito nell'articolo 1, paragrafo 4, della direttiva 91/689/CEE del Consiglio, del 12 dicembre 1991, relativa ai rifiuti pericolosi".

Quest'ultimo risultato apre interessanti prospettive future sull'affidabilità del metodo di estrazione automatica di terminologia qui adottato come strumento a supporto dell'individuazione delle **definizioni** all'interno di un corpus di testi legislativi. Centrali in un'ottica di accesso al contenuto giuridico visto dai giuristi come un processo di interpretazione del testo, le definizioni sono altrettanto centrali in una prospettiva di rappresentazione ed estrazione della conoscenza contenuta in una collezione documentale. Come messo in evidenza da Walter (2009), il loro riconoscimento è fondamentale per l'annotazione semantica di un testo giuridico[19] oltre che come strategia

[18]Il Regolamento è parte del corpus AMBnorm(Europa).
[19]Vedi Paragrafo 2.3.2.2.

finalizzata alla costruzione di ontologie giuridiche a partire da testi.

5.3 La "collocazione del lessico nel contesto degli enunciati": la sintassi come punto di partenza per l'annotazione semantica

Il secondo passo ritenuto fondamentale per accedere con successo al contenuto informativo dei testi giuridici è quello rappresentato dal riconoscimento delle relazioni semantico–combinatorie che legano i termini all'interno del contesto degli enunciati. Un tale processo è qui inteso come un processo di **annotazione semantica**, il quale fornendo una rappresentazione formale, a livello sintagmatico, del significato di alcune delle più significative unità lessicali presenti nei testi permette di rendere esplicito l'intero contenuto proposizionale degli enunciati. Fondandosi cioè sulla ricostruzione dei rapporti sintagmatici esistenti tra termini nel contesto di un periodo, la metodologia di annotazione semantica qui messa a punto mira a individuare tutti gli elementi informativi presenti nel testo e indispensabili per rendere pienamente esplicito il contenuto del testo stesso.

L'orizzonte di riferimento è quello degli approcci alla rappresentazione strutturata del significato sviluppati nell'ambito degli studi di semantica lessicale. In particolare, si è qui scelto di adottare i principi della 'Frame Semantics Theory' intesa, nell'interpretazione di Charles Fillmore, come una 'semantics of understanding'[20]. Ai fini di uno studio linguistico empirico di testi giuridici articolato su più livelli di analisi testuale, come quello qui realizzato, essa si configura infatti come uno strumento teorico indispensabile.

Avendo come obiettivo lo studio del linguaggio naturale attraverso la sua concreta realizzazione nel testo, Fillmore (1985) suggerisce la necessità di ripensare a cosa si debba intendere, in particolare, per "language–internal *semantic representation* of a sentence". Egli ritiene che per rappresentare il significato di un periodo sia necessario individuare prima di tutto quali

[20]Fillmore (1985) definisce la 'semantics of understanding' come la teoria che "takes as its assignment that of providing a general account of the relation between linguistic texts, the contexts in which they are instanced, and the process and products of their interpretation". In questo senso, essa si contrappone alla 'semantics of truth', dal momento che quest'ultima "by contrast, begins by assuming that its goal is to characterize the conditions under which individual utterances of a given language can be said to be true".

Posizione nella lista estratta	Termine	Tipologia
Unità poliremiatiche estratte sulla base del C–NC Value		
1	*parlamento europeo*	tecnico–giuridico
2	*autorità competente*	tecnico–giuridico
3	*valore limite*	ambientale
4	*valore limite di emissione*	ambientale
5	*limite di emissione*	ambientale
6	*presente regolamento*	tecnico–giuridico
7	*gas ad effetto serra*	ambientale
8	*immissione sul mercato*	tecnico–giuridico
9	*applicazione della presente*	errore
10	*riduzione delle emissioni*	ambientale
11	*qualità dell'aria*	ambientale
12	*stato membro*	tecnico–giuridico
13	*disposizione dell'articolo*	tecnico–giuridico
14	*disposizione del presente*	errore
15	*sostanza pericolosa*	ambientale
16	*gas ad effetto*	errore
26	*effetto serra*	ambientale
39	*giorno successivo alla pubblicazione*	comune
41	*ravvicinamento delle disposizioni legislative*	tecnico–giuridico
292	*diritto nazionale*	tecnico–giuridico
Unità poliremiatiche estratte dopo il contrasto con PAROLE		
1	*parlamento europeo*	tecnico–giuridico
2	*presente regolamento*	tecnico–giuridico
3	*valore limite*	ambientale
4	*valore limite di emissione*	ambientale
5	*immissione sul mercato*	tecnico–giuridico
6	*destinatario della presente*	errore
7	*riduzione delle emissioni*	ambientale
8	*gas ad effetto serra*	ambientale
9	*stato membro*	tecnico–giuridico
10	*limite di emissione*	ambientale
11	*parere del comitato*	tecnico–giuridico
12	*sostanza pericolosa*	ambientale
13	*organico persistente*	errore
14	*aria ambiente*	ambientale
15	*applicazione della presente direttiva*	tecnico–giuridico
16	*rifiuti di imballaggio*	ambientale
37	*effetto serra*	ambientale
22	*applicazione della presente*	errore
40	*ravvicinamento delle disposizioni legislative*	tecnico–giuridico
228	*giorno successivo alla pubblicazione*	comune
Unità poliremiatiche estratte dopo il contrasto con CONS		
1	*valore limite*	ambientale
2	*sostanza pericolosa*	ambientale
3	*salute umana*	ambientale
4	*sviluppo sostenibile*	ambientale
5	*principio attivo*	ambientale
6	*inquinamento atmosferico*	ambientale
7	*limite di emissione*	ambientale
8	*effetto serra*	ambientale
9	*rifiuto pericoloso*	ambientale
10	*valore limite di emissione*	ambientale
288	*disposizione legislativa*	tecnico–giuridico
289	*norma nazionale*	tecnico–giuridico
290	*disposizione della presente direttiva*	tecnico–giuridico
292	*livello di protezione*	tecnico–giuridico
294	*diritto interno*	tecnico–giuridico
295	*diritto nazionale*	tecnico–giuridico
296	*ravvicinamento delle disposizioni legislative*	tecnico–giuridico
298	*testo della disposizione essenziale del diritto*	tecnico–giuridico
299	*disposizione essenziale del diritto interno*	tecnico–giuridico
300	*disposizione nazionale*	tecnico–giuridico

Tabella 5.1: Alcuni dei termini poliremiatici estratti sulla base del C–NC Value, dopo il contrasto con PAROLE e con CONS.

siano gli elementi minimi che lo descrivono e come essi siano linguisticamente realizzati nel testo. Di conseguenza, "a language–internal semantic parsing of a sentence must be seen as merely a display of the lexical, grammatical and semantic material of the sentence".

Il processo di accesso al contenuto informativo di un testo si configura pertanto come un processo di progressiva identificazione di **tutti** gli elementi conoscitivi rintracciabili all'interno del testo stesso. Chiarisce Fillmore (1985): "I view the process of interpreting a linguistic text as that of giving it a maximally rich interpretation, an interpretation which draws everything out of the text that it can". Tale identificazione si concretizza allora in un'analisi del testo articolata su più fasi di analisi tra loro strettamente collegate e complementari.

È qui d'interesse far notare come una tale prospettiva di indagine incrementale del testo abbia somiglianze con quella prospettata dalla scuola italiana di filosofia analitica del diritto. Come illustrato da Jori e Pintore (1995, p. 212), l'attività di interpretazione di un testo giuridico non è altro che un'attività di analisi linguistica del testo la quale consiste in "una attenta considerazione della struttura sintattica e grammaticale; una comprensione del suo lessico; una collocazione di questo nel contesto degli enunciati".

La possibilità dunque di rendere esplicito il contenuto informativo di testi giuridici si concretizza metodologicamente in questo lavoro in un processo di annotazione stratificata del testo, durante la quale, in una prima fase di annotazione morfosintattica e sintattica del testo giuridico, viene resa esplicita l'informazione linguistica in esso contenuta e, in una seconda fase, l'annotazione semantica permette di identificare gli elementi informativi necessari a rendere esplicito il significato degli enunciati in esso contenuti.

In questa prospettiva, l'annotazione sintattica si configura come punto di partenza privilegiato per l'analisi della dimensione semantica di un testo giuridico. L'attenzione dedicata in questo lavoro a questo livello di descrizione linguistica è giustificata da più punti di vista: *i)* dagli studi linguistici sulla lingua del diritto, *ii)* da chi ha visto nell'analisi del linguaggio il carattere fondante l'intero processo di interpretazione del discorso giuridico, *iii)* da chi, con finalità applicative, assume il risultato dell'annotazione sintattica del testo come base di una fase di annotazione semantica, a sua volta punto di partenza per compiti di gestione automatica del contenuto.

Come messo in evidenza nel Paragrafo 2.1, le indagini di Rovere (2005) sui rapporti tra significato e quadro valenziale di verbi comuni e tecnici che occorrono in corpora di testi giuridici si collocano proprio in questa prospet-

tiva. Mettendo in relazione i componenti del quadro valenziale di un verbo (parte del suo "frame" sintattico), così come sono realizzati nel testo, con i corrispondenti valori semantici di dominio, "a differenza degli approcci spesso intuitivi alla valenza pragmatica dei verbi [...], è possibile fondare il frame su categorie tecniche identificabili con precisione" (Rovere, 2005, p. 162)[21]. È possibile cioè dare una giustificazione empirica alla "configurazione tecnica degli argomenti" di un verbo presente in un testo giuridico.

Un tale approccio riecheggia la prospettiva di indagine della 'Frame Semantics Theory' di Fillmore. È l'idea, esposta da Fillmore e Atkins (1992), in base alla quale un compito di rappresentazione esplicita del significato di un'unità lessicale consiste in un processo di "*valence description*, a description that specifies, in both semantic and syntactic terms, what the expression requires of its constituents and its context, and what it contributes to the structure that contain it". Come ricorda Hanks (2002), interpretando il pensiero di Fillmore e Atkins: "First, let's get the syntactic structure clear [...] Then let's relate that structure [...] to the whole conceptual framework within which the word exists".

Quando Scarpelli (1969) ricorda che nella sua attività interpretativa il giurista è impegnato in un processo continuo in cui "ad ogni passo egli deve [...] riconoscere, costruire o ricostruire relazioni semantiche, e sintattiche e pragmatiche", egli allude all'importanza, riconosciuta nell'ambito degli studi della scuola italiana di filosofia analitica del diritto, di mettere in relazione il significato con il concreto uso linguistico.

È qui inoltre d'interesse ricordare come da più parti all'interno della comunità di ricerca in AI&Law sia riconosciuto il fatto che il processo di annotazione linguistica del testo giuridico sia il punto di partenza per successivi compiti di gestione della conoscenza di dominio in esso contenuta. Ne sono in particolare una dimostrazione gli studi basati sull'uso di strumenti di Trattamento Automatico del Linguaggio finalizzati all'annotazione semantica di testi giuridici. Come messo in luce nel Paragrafo 2.3.2.2, essi sono infatti accomunati dal ricorso ad una fase di annotazione sintattica del testo come punto di partenza per individuare elementi semanticamente rilevanti nella collezione documentale di riferimento.

Tuttavia, come ricorda Hanks (2000), "there is no direct route from the corpus to the meaning". Ingrediente fondamentale in un processo di anno-

[21]Nota che in questo contesto Rovere con il termine "frame" allude alla struttura sintattico–valenziale di un verbo.

tazione semantica è infatti un modello di riferimento rispetto al quale organizzare gli elementi semanticamente rilevanti individuati in un corpus. Allo scopo di rendere esplicito il contenuto informativo presente in una collezione di testi, è dunque necessario disporre di "modelli di rappresentazione, atti a dar conto di tutti i fatti linguistici presenti nel corpus"[22], di modelli che permettano di descriverne il contenuto semantico–lessicale sulla base di una serie di principi organizzativi.

È questo il tema affrontato nei due successivi capitoli, dove sono prima di tutto descritte (nel Capitolo 6) le motivazioni che hanno portato a scegliere la 'Frame Semantics Theory' di Fillmore come teoria di riferimento del modello di rappresentazione del significato adottato nell'annotazione semantica di testi giuridici. Nel Capitolo 7 sono inoltre riportati i risultati di un caso di studio finalizzato ad illustrare come la metodologia di annotazione semantica messa a punto in questo lavoro possa essere concretamente applicata per rendere esplicito il contenuto deontico di enunciati normativi.

[22]È il punto di vista assunto in particolare di Rovere (2005), dove l'autore sottolinea la necessità di modelli organizzativi degli elementi linguistici rintracciati in un corpus.

Capitolo 6

Un modello per l'annotazione semantica di testi giuridici

Questo capitolo è dedicato a mettere in luce come la 'Frame Semantics Theory' elaborata da Charles Fillmore e i principi organizzativi sottesi al progetto FrameNet siano particolarmente adatti a offrire un'esaustiva rappresentazione del contenuto semantico–lessicale di testi giuridici.

Come discusso nei paragrafi che seguono, tale convinzione parte da due presupposti fondamentali. In primo luogo, essa è fondata sull'idea che il modello FrameNet e i principi teorici che lo governano possano essere utilizzati in un processo di annotazione semantica di corpora giuridici finalizzato a renderne esplicito il contenuto informativo. Grazie ad un principio di organizzazione sintagmatica e 'stratificata' del significato, tale modello offre gli strumenti adeguati per rappresentare le proprietà semantico–combinatorie delle parole a partire dalle strutture sintattiche nelle quali esse ricorrono nel testo. In tal senso, esso si configura come un modello complementare al modello sino ad oggi utilizzato per l'organizzazione della conoscenza semantico–lessicale giuridica, modello basato su principi di organizzazione paradigmatica del significato.

In secondo luogo, questo capitolo trae ispirazione dall'intuizione che ogni modello di rappresentazione del significato progettato per organizzare e strutturare il contenuto di corpora di lingua comune possa essere adattato con successo per rappresentare quello di corpora di dominio. È quanto è stato recentemente dimostrato da Dolbey (2009). Egli per primo ha infatti mostrato come l'annotazione semantica di articoli scientifici in materia di biologia molecolare basata sui principi organizzativi di FrameNet sia utile per diverse

comunità di ricerca, *i)* offrendo nuove prospettive per lo studio linguistico del linguaggio biomedico, *ii)* ponendo le basi per lo sviluppo di un FrameNet di dominio e *iii)* fornendo una soluzione al ben noto problema di trovare un collegamento tra l'informazione semantico–lessicale contenuta in corpora di dominio e la conoscenza di dominio organizzata in sistemi di organizzazione della conoscenza (ontologie) costruiti in modo 'astratto', unicamente a partire cioè dalla conoscenza di dominio senza fare riferimento alla realizzazione linguistica nel testo.

Lo scopo è qui quello di illustrare le potenzialità di FrameNet per il dominio giuridico sia rispetto ad un modello di rappresentazione del significato, basato su un'organizzazione **paradigmatica** dello spazio semantico–lessicale di una parola, sia rispetto agli altri progetti lessicografici oggi avviati e finalizzati alla rappresentazione del significato a livello **sintagmatico**.

Per questo, i principi organizzativi di FrameNet sono qui messi a confronto *i)* con quelli di WordNet, il principale lessico computazionale in cui i significati delle parole sono gerarchicamente organizzati in una rete di relazioni semantiche (Paragrafo 6.2) e *ii)* con quelli degli altri progetti che permettono di rendere esplicito il significato di una parola sulla base delle sue proprietà semantico–combinatorie (Paragrafo 6.3).

La rassegna, condotta nel Paragrafo 6.4, dei vari usi e specializzazioni di dominio che dei modelli di rappresentazione del significato sono stati fatti è finalizzata a mettere in luce come sino ad oggi poca attenzione sia stata dedicata al dominio giuridico. L'unica eccezione è rappresentata da JurWordNet, specializzazione per la lingua italiana giuridica del modello WordNet.

Le discussioni circa le potenzialità dell'uso di FrameNet come modello di annotazione semantica (Paragrafo 6.5) prendono pertanto le mosse da questa articolata serie di confronti.

6.1 Il modello FrameNet di rappresentazione sintagmatica del significato

Avviato agli inizi degli anni '90 nell'ambito delle attività di ricerca svolte da Charles Fillmore e dai suoi collaboratori presso l'Università della California (Berkeley), il progetto FrameNet[1] (Baker et al., 1998) è espressamente finalizzato alla costruzione *i)* di un lessico computazionale basato su attesta-

[1] http://framenet.icsi.berkeley.edu/

zioni d'uso in un corpus testuale e *ii)* di un corpus annotato con informazioni relative alle proprietà semantico–combinatorie delle parole nel testo[2]. Inizialmente la collezione di frasi annotate è stata selezionata a partire dal British National Corpus[3]. Il processo di annotazione è tutt'ora in corso e si è esteso anche all'annotazione di altri corpora testuali, tra i quali anche l'American National Corpus[4].

È qui d'interesse sottolineare come dunque il progetto abbia rilevanza sia in ambito lessicografico, come lessico computazionale in grado di superare alcuni dei limiti dei dizionari tradizionali (Fillmore e Atkins, 1994), sia in ambito di rappresentazione formale del significato, come modello di rappresentazione del contenuto semantico di un testo usato per realizzare compiti di gestione automatica dell'informazione basati su metodi e strumenti di Trattamento Automatico del Linguaggio (Lowe et al., 1997).

6.1.1 I fondamenti teorici della Frame Semantics Theory

FrameNet si presenta come la realizzazione pratica dei principi teorici della 'Frame Semantics Theory' di Fillmore[5], intesa come "the study of how, as part of our knowledge of the language, we associate linguistic forms (words, fixed phrases, grammatical patterns) with the cognitive structures – the frames – which largely determine the process (and the result) of interpreting those forms" (Fillmore e Baker, 2010, p. 314).

Non è tra gli obiettivi di questo lavoro quello di discutere in modo esaustivo i principi della 'Frame Semantics'. È piuttosto qui d'interesse mettere in evidenza quegli aspetti che rendono questa teoria uno strumento particolarmente adatto per essere usato come chiave di accesso alla semantica lessicale della lingua del diritto. In questo senso è fondamentale la prospettiva **olistica** e **descrittiva** con cui si guarda al significato di una parola e, in particolare, ai processi di (ri)costruzione del significato.

[2] "The aim is to document the range of semantic and syntactic combinatory possibilities – valences – of each word in each of its senses, through computer–assisted annotation of example sentences and automatic tabulation and display of the annotation results." (Ruppenhofer et al., 2010, p. 5).

[3] http://www.natcorp.ox.ac.uk/

[4] http://www.americannationalcorpus.org/

[5] Originariamente esposta da Fillmore (1985).

L'aspetto olistico è legato all'unità minima di rappresentazione del significato considerata: il 'frame'[6], definito da Fillmore (1982, p. 111) come "any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits". In questo senso, il frame è uno strumento concettuale che permette di enfatizzare l'aspetto pragmatico–cognitivo del processo di (ri)costruzione del significato di una parola, riconducendolo ad un processo interpretativo durante il quale ogni elemento che contribuisce alla rappresentazione del spazio semantico–lessicale viene ricondotto a un "'prototype' rather than [...] a genuine body of assumptions about what the world is like" (Fillmore, 1982, p. 118).

Il frame è la rappresentazione schematica di un dato bagaglio di conoscenze relative ad una situazione–tipo descritta nei suoi singoli componenti. In altri termini, esso costituisce il modello formale che consente una descrizione dei processi cognitivi di comprensione attivati in uno scambio comunicativo attraverso la loro rappresentazione astratta e prototipica[7].

Un secondo aspetto degno di rilievo della 'Frame Semantics' consiste nell'usare uno strumento concettuale ampiamente utilizzato in letteratura, come il frame, applicandolo per la prima volta anche all'analisi testuale. L'idea per cui il frame sia "a useful tool in lexical semantics, in the semantics of grammar, and in text semantics" (Fillmore, 1985, p. 222) permette infatti di rendere esplicito il collegamento tra 'conceptual frame' e descrizione linguistica di un testo.

È questo il motivo per cui la 'Frame Semantics' rappresenta l'orizzonte teorico di riferimento sia *i)* di compiti di descrizione del contenuto proposizionale di una frase, di attività cioè di annotazione semantica finalizzate a rendere esplicito il rapporto tra realizzazione linguistica (grammaticale) e organizzazione sintagmatica dei principali componenti informativi di un testo sia *ii)* di compiti lessicografici di rappresentazione dello spazio semantico–lessicale di una parola, secondo un approccio all'organizzazione del significato

[6] "I intend the word 'frame' as used here to be a general cover term for the set of concepts variously known, in the literature on natural language understanding, as 'schema', 'script', 'scenario', 'ideational scafolding', 'cognitive model', or 'folk theory' " (Fillmore, 1982, p. 111).

[7] Come già precedentemente discusso nel Paragrafo 5.3, questa prospettiva olistica sul significato è esplicitamente fondata sui presupposti della "U–semantic theory" (o "semantics of understanding"), la teoria per la quale "we can know the meanings of the individual words only by first understanding the factual basis for the relationship which they identify" (Fillmore, 1985, p. 224).

focalizzata sulla descrizione di "what the expression requires of its constituents and its context, and what it contributes to the structures that contain it" (Fillmore e Atkins, 1992, p. 78), piuttosto che sull'elenco di significati.

È chiaro allora come un tale approccio, sottolineando il ruolo centrale del contesto d'uso delle parole e delle 'regole' prototipiche che ne governano l'uso, sia un modello formale particolarmente adatto per un compito di semiotica giuridica intesa come semiotica linguistica[8]. È di fatto in linea con la concezione dei rappresentanti della scuola analitica italiana di filosofia del diritto per i quali "il significato di una parola non è qualcosa che sia intrinsecamente e definitivamente legato ad essa" (Scarpelli, 1976b), ma è determinato dalle regole d'uso stabilite in un universo concettuale condiviso.

È qui inoltre d'interesse sottolineare come Fillmore (1982) stesso, mettendo in evidenza i vantaggi che la 'Frame Semantics' offre per affrontare con successo alcune dibattute questioni di semantica lessicale, proponga esempi tratti proprio dal dominio giuridico. Tale dominio infatti bene esemplifica le opportunità di un tale punto di vista empirico/descrittivo sulle dinamiche del significato. I casi discussi, in cui la 'Frame Semantics' aiuta a spiegare questioni complesse, sono i seguenti tre:

- il caso di uso semanticamente errato nel linguaggio comune di un termine impiegato per descrivere uno stato di cose esplicitamente regolato dalla dottrina giuridica, per riferirsi al quale bisognerebbe usare un termine specifico. L'esempio è quello del termine *culprit* erroneamente usato nel linguaggio comune al posto del (giuridicamente) corretto *suspect* per indicare una persona solo sospettata di aver commesso un reato, ma di cui non sia ancora stata provata in giudizio la colpevolezza. Tale scambio lessicale è spiegato nei termini della 'Frame Semantics' come un caso in cui "the links between words and their frames are changed, but the underlying schematization remains unchanged" (Fillmore, 1982, pp. 126-127);

- il fraintendimento comunicativo che si viene a creare nel caso di uso di parole che hanno significati diversi nella lingua del diritto e nel linguaggio comune. Il caso portato come esempio è quello della coppia di opposti *innocent/guilty*. Nel linguaggio comune una persona è innocente o colpevole se non ha o ha commesso un reato, mentre in base alla dottrina giuridica una persona continua a rimanere innocente finché

[8]Vedi il Paragrafo 2.2.

la sua colpevolezza non viene provata da un tribunale. Sono questi i casi in cui "new framings need to be constructed for familiar words" (Fillmore, 1982, pp. 127-128);

- i casi di estensioni di significato dovute all'uso specialistico di un termine. L'esempio è quello del termine *oral agreement* che in base alla dottrina giuridica ha significato (e validità legale) di contratto, anche se non è scritto e firmato, a differenza invece del significato generico che ha nel linguaggio comune. Un caso simile è spiegabile come un caso di cambio "between general and special–purpose framings of words" (Fillmore, 1982, pp. 128-129).

Inoltre, altri due aspetti della 'Frame Semantics Theory' sono ritenuti centrali in questo lavoro. In primo luogo, uno dei suoi principi cardine, l'idea cioè che "what happens when one comprehends a text is that one mentally creates a kind of world" (Fillmore, 1977, p. 61). Basato sulla concezione che ogni processo di comprensione del testo non sia altro che "the process of interpreting language in context" (Fillmore, 1977, p. 64), un compito di annotazione semantica di testi giuridici basata sui principi teorici della 'Frame Semantics' e sui principi organizzativi di FrameNet può essere visto come un processo finalizzato alla verifica di come uno o più frame(s) si istanziano correttamente nel testo, consentendo in questo modo la piena comprensione da parte del lettore del contenuto proposizionale[9].

Infine, di fondamentale importanza per una completa descrizione dell'informazione semantico–lessicale contenuta in testi giuridici è la possibilità offerta dalla 'Frame Semantics' di guardare ad una data situazione–tipo sia (in generale) nella sua complessità sia (in particolare) da diversi punti prospettici di osservazione. Ciò è permesso dai due livelli di descrizione previsti: uno che restituisce "a fairly complete understanding of the nature of the total transaction or activity" e uno che fornisce "a particular perspectival anchoring among the entities involved in the activity" (Fillmore, 1977, p. 59). Come discusso nel Paragrafo 6.5.2 ed esemplificato nel Capitolo 7, questa visione consente di avere una descrizione dei principali concetti giuridici più sfaccettata di quella monolitica fornita dalle ontologie giuridiche formali.

[9] È la prospettiva di analisi sperimentata con successo da Rathert (2006) e descritta nel Paragrafo 6.4.3.

6.1.2 I principi e gli elementi organizzativi di FrameNet

L'unità minima di rappresentazione del significato in FrameNet è il frame. A partire dal database consultabile on–line, l'informazione semantico–lessicale è organizzata come segue. Ogni frame contiene:

- una **definizione** che descrive la situazione–tipo descritta dal frame; ad esempio, del frame DEPARTING viene data la seguente descrizione "An object (the Theme) moves away from a Source. The Source may be expressed or it may be understood from context, but its existence is always implied by the departing word itself";

- una **lista di ruoli semantici** (i 'Frame Elements', d'ora in avanti FEs) che descrivono il ruolo giocato dai partecipanti alla situazione–tipo, un ruolo a cui è associato un nome che ne caratterizza la funzione 'specifica' svolta in ogni singolo frame, invece che un ruolo tematico 'astratto' del tipo 'agente', 'paziente', ecc..., e una descrizione. I FEs sono divisi in partecipanti 'Core' e 'Non–Core', distinzione stabilita unicamente su base semantica a partire dalla centralità che un ruolo gioca nella descrizione di una situazione–tipo (Ruppenhofer et al., 2010, pp. 19–21). Così, ad esempio, il frame DEPARTING comprende *i)* i 'Core' FEs Source[10] e Theme[11] e *ii)* una lunga lista di 'Non–Core' FEs, quali Circumstances, Goal, Distance, Path, Mode_of_transportation, ecc...;

- una **caratterizzazione ontologica** ('Semantic Type', d'ora in avanti ST) di ogni FE, la restrizione di selezione semantica (ontologica) delle istanze di ogni ruolo semantico nelle annotazioni testuali. Tutti i FEs a cui è associato lo stesso nome identificativo hanno lo stesso ST. Così, ad esempio, tutti i FEs Source in FrameNet hanno come ST associato Location, i FEs Theme sono Physical_object, ecc...

  L'obiettivo è quello di fornire informazioni semantiche aggiuntive che non sono contenute nella struttura gerarchica di FrameNet (Ruppenhofer et al., 2010, pp. 79–80). A questo scopo, sono stati previsti 40 STs organizzati in modo gerarchico e parzialmente collegati a WordNet;

[10] Il FE in FrameNet è definito come "any constituent that expresses the initial position of the Theme, before the change of location".

[11] Il FE è definito come "the object which moves. It may be an entity which moves under its own power, but it need not be".

- una **serie di relazioni di ereditarietà** ('frame–to–frame relations'), quali ad esempio Inheritance, Using, Perspective On, ecc..., che legano tra loro i frames contenuti nel database di FrameNet, definendone così la struttura 'a rete' (Ruppenhofer et al., 2010, pp. 73–79). Si tratta di relazioni dirette (asimmetriche) che mettono in collegamento due frames, un 'Super frame' più astratto (il 'padre' nella relazione di ereditarietà) e un 'Sub frame' meno astratto (il 'figlio').

  Come si può vedere nella Figura 6.1[12], il frame DEPARTING, è legato, ad esempio, da una relazione Using[13] al frame MOTION[14], di cui è il 'Sub frame', e al frame DISEMBARKING[15], di cui è il 'Super frame'. Ciò implica che nel primo caso il frame DEPARTING 'usa' alcuni dei FEs del 'Super frame' MOTION; nel secondo caso, invece, alcuni dei FEs di DEPARTING sono 'usati' dal 'Sub frame' DISEMBARKING. Il collegamento tra i due frames è a livello dei singoli FEs; così, ad esempio, i FEs Source e Theme del frame DEPARTING corrispondono (sono 'usati') rispettivamente al FE Vehicle e Traveller del 'Sub frame' DISEMBARKING; il FE Place di DEPARTING 'usa' il FE Area del 'Super frame' MOTION;

- una **lista di unità lessicali** ('lexical units', d'ora in avanti LUs), appartenenti a categorie morfosintattiche diverse, che rimandano ('evocano') un determinato frame. Ad esempio, il frame DEPARTING è evocato dalle seguenti LUs: *decamp.v, depart.v, departure.n, disappear.v, disappearance.n, emerge.v, escape.n, escape.v, exit.n, exit.v, exodus.n,* ecc... È infatti sulla base della struttura argomentale (o valenziale) delle LUs

[12]Questa figura è stata realizzata grazie al FrameGrapher, lo strumento di rappresentazione grafica delle relazioni 'frame–to–frame' utilizzabile on–line alla pagina http://framenet.icsi.berkeley.edu/FrameGrapher/

[13]La relazione Using viene stabilita nel caso particolare in cui "a particular frame makes reference in a very general kind of way to the structure of a more abstract, schematic frame"; per questo motivo "is used almost exclusively for cases in which a part of the scene evoked by the Child refers to the Parent frame" (Ruppenhofer et al., 2010, pp. 78)

[14]Il frame MOTION descrive una situazione nella quale "Some entity (Theme) starts out in one place (Source) and ends up in some other place (Goal), having covered some space between the two (Path). Alternatively, the Area or Directionin which the Theme moves or the Distance of the movement may be mentioned".

[15]Il frame DISEMBARKING descrive una situazione nella quale "A Traveller leaves from or dismounts a Vehicle".

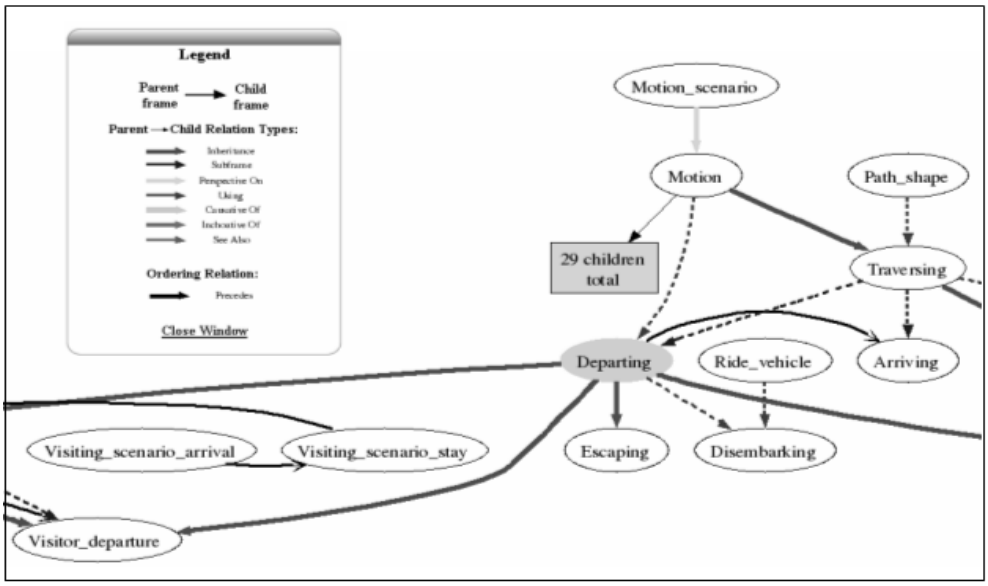


Figura 6.1: Rete di relazioni 'frame–to–frame' in cui è inserito il frame DEPARTING.

individuata in un periodo in fase di annotazione che sono rintracciate le possibili realizzazioni lessicali dei FEs di un frame.

Sulla base di questi elementi, FrameNet si configura come un database sia di informazioni lessicografiche sia di annotazioni testuali. Per ogni LU è infatti riportata *i)* un'entrata lessicale che contiene, oltre alla descrizione del frame evocato, la lista di FEs con tutte le possibili corrispondenti realizzazioni sintattiche ('valence patterns') nelle frasi annotate e *ii)* la lista di periodi del British National Corpus annotati.

I periodi sono annotati come negli esempi che seguono, dove sono state riportate alcune delle realizzazioni del frame DEPARTING evocato dal verbo *to depart* e dal sostantivo *escape*[16]:

(a) On the eve of World War II, [both James and Eric Williams $_{Theme}$] **departed** [England $_{Source}$] [for the United States $_{Goal}$].

[16]La LU che evoca il frame DEPARTING è evidenziata in grassetto; i FEs sono riportati a pedice.

(b) So [Rodrigo $_{Theme}$] **departed** [from the King $_{Source}$], and took his spouse with him to the house of his mother, and gave her to his mother's keeping.

(c) [The swifts $_{Theme}$] will **depart** [in the first week of August $_{Time}$]. [DNI $_{Source}$]

(d) Second, fundamentalism is a ghetto–like **escape** [from the world $_{Source}$]. [CNI $_{Theme}$]

(e) [The men $_{Theme}$] [made $_{Supp}$] [their $_{Theme}$] **escape** [in Munn 's car $_{Mode_of_transportation}$], which he had reported stolen earlier that day, but witnesses contacted police with the registration number. [DNI $_{Source}$]

(f) This was not so much an **escape** as a therapy. [INI $_{Source}$] [CNI $_{Theme}$]

Come mostra la Figura 6.2[17], dove è stata schematizzata l'annotazione di (a), in FrameNet viene resa esplicita l'informazione relativa *i)* al tipo di costituente morfosintattico dei singoli partecipanti (FEs) alla situazione evocata, *ii)* alla funzione grammaticale ricoperta dai FEs nel periodo e *iii)* al ruolo semantico giocato dai partecipanti alla situazione evocata[18].

Pertanto, grazie alla rappresentazione esplicita di come il contenuto semantico di un'unità lessicale è linguisticamente (morfosintatticamente e sintatticamente) istanziato in un testo, FrameNet si presenta come uno strumento di descrizione lessicografica più 'potente' e flessibile dei modi tradizionali, fornendo la possibilità "of seeing a single 'sense' (i.e a single underlying schematization) realized in different syntactic forms" (Fillmore e Atkins, 1994, p. 370).

Come dimostrano gli esempi di annotazione riportati sopra, ad esempio, una tale rappresentazione del significato permette di catturare il fatto che 'la posizione iniziale da cui un'entità si muove per spostarsi' (FE Source) può essere espressa in vari modi: *i)* come sintagma nominale che svolge la funzione di 'oggetto' del verbo *to depart*, esempio (a); ma anche *ii)* come

[17]L'annotazione delle categorie morfosintattiche e delle funzioni grammaticali corrispondenti è realizzata sulla base dello schema di annotazione adottato in FrameNet.

[18]In questo caso, l'annotazione permette di rendere esplicito, ad esempio, il fatto che la meta del viaggio dei Williams, FE Goal, è realizzata come un sintagma preposizionale introdotto dalla preposizione *for* (PP[for]) e legato da una relazione sintattico–funzionale di 'dependent' alla LU *to depart*.

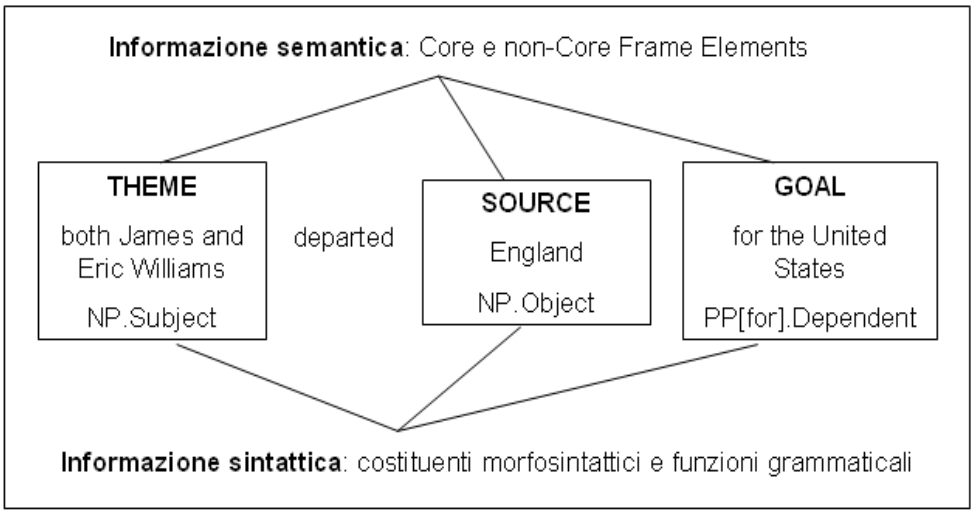


Figura 6.2: Annotazione di *to depart*.DEPARTING.

suo 'complemento' realizzato da un sintagma preposizionale introdotto dalla preposizione *from*, esempio (b); o ancora *iii)* può essere lasciato implicito, non essere cioè istanziato affatto sebbene esso costituisca materiale semanticamente rilevante per la piena ricostruzione del contenuto proposizionale di una frase, esempio (c).

È quest'ultimo un esempio particolarmente significativo del contributo innovativo di FrameNet alla lessicografia, come discusso da Atkins et al. (2003a). Il fatto di rendere conto anche delle condizioni di omissibilità di partecipanti ad un frame consente infatti di fornire una descrizione anche degli usi idiosincratici di un'unità lessicale. A questo scopo in FrameNet sono previsti tre tipi di 'non istanziazione' di FEs (Ruppenhofer et al., 2010, pp. 24–26):

– 'Definite Null Instantiation' (DNI), quando il materiale semantico–lessicale omesso è desumibile dal contesto, è dunque 'definito' in maniera anaforica come nel caso dell'esempio (c);

– 'Indefinite Null Instantiation' (INI), i casi di omissione 'esistenziale' di partecipanti ad un frame, che non sono espressi né sono deducibili dal contesto, come nel caso dell'esempio (f) dove è chiaro che si sta

facendo riferimento ad una fuga 'da qualcosa' ma non è in nessun modo definibile la sorgente di tale fuga;

– 'Constructional Null Instantiation' (CNI), determinata dalla struttura grammaticale nella quale una determinata LU occorre, come nel caso dell'omissione di soggetti di forme imperative, di agenti di verbi passivi, di strutture a soggetto controllato o ancora, come nel caso degli esempi (d) e (f), quando la particolare struttura del periodo ammette che 'l'oggetto fisico che fugge' (FE Theme) sia omesso.

Una seconda potenzialità di FrameNet come modello di descrizione semantico–lessicale riguarda il modo in cui vengono trattate le costruzioni supporto. È il caso, ad esempio, del periodo (e) nel quale il riferimento al frame DEPARTING è attivato dal sostantivo *escape* e non dal verbo *to make* (annotato come $_{Supp}$). Ciò permette di ampliare la prospettiva lessicografica non più ristretta al contributo semantico di una singola unità lessicale predicativa, ma estesa al contesto sintattico nel quale essa è inserita.

È in questo modo anche possibile considerare i diversi aspetti lessicograficamente rilevanti delle costruzioni verbo supporto (Ruppenhofer et al., 2010, pp. 31–38), come ad esempio il fatto che *i)* significati diversi di un sostantivo posso essere selezionati da verbi supporto diversi (es. ***have*** *an argument* evoca il frame QUARRELING, mentre ***make*** *the argument* evoca il frame REASONING), *ii)* verbi supporto diversi possono presuppore partecipanti ad un evento coinvolti con ruoli diversi (es. ***perform*** *an operation* vs ***undergo*** *an operation*), *iii)* verbi supporto diversi possono fare riferimento a fasi diverse di un evento complesso (es. ***make*** *a promise* vs ***keep*** *a promise*).

Il trattamento delle costruzioni supporto in FrameNet riguarda anche i casi di preposizioni supporto (Ruppenhofer et al., 2010, pp. 38–39). Come per i verbi supporto, anche in questo caso è il sostantivo l'elemento semanticamente determinante che seleziona la preposizione con cui occorre. È il caso, ad esempio, della preposizione *in* che in unione al sostantivo *contravention* evoca il frame COMPLIANCE, come mostra il seguente periodo annotato:

- [*This traffic* $_{State_of_Affairs}$] *was the object of UN toleration in spite of being theoretically* [*in* $_{Supp}$] ***contravention*** [*of UN sanctions against Iraq* $_{Norm}$].

È qui infine d'interesse ricordare che, in linea con la duplice finalità del progetto, quella cioè di costruire un lessico computazionale basato su un cor-

pus annotato con informazione semantica, nell'ambito del progetto FrameNet sono state messe a punto due modalità di annotazione testuale:

– una di tipo lessicografico ('lexicographic annotation'), limitata a quei periodi che contengono LUs precedentemente selezionate perché di interesse lessicografico. In questo caso l'obiettivo è quello di raccogliere una serie di attestazioni d'uso reale che testimonino le diverse possibilità combinatorie (sintattiche e semantiche) di una LU in tutti i suoi sensi, arricchendo con evidenza testuale le entrate lessicali del lessico FrameNet;

– una di tipo 'continuo' ('full–text annotation'), che prevede l'annotazione di tutte le LUs presenti in un periodo in grado di evocare un frame. Al centro delle più recenti attività del progetto FrameNet, l'obiettivo di quest'ultima modalità è quello di dimostrare come il modello di annotazione FrameNet sia un utile strumento di comprensione testuale.

6.1.3 Gli usi di FrameNet

Allo scopo di collocare il presente lavoro nel contesto delle numerose attività di ricerca che a livello internazionale sono basate su FrameNet, in quanto segue sono passati in rassegna i diversi usi che sino ad oggi sono stati fatti (e/o sono tutt'ora in corso) del modello originario di Berkeley[19].

I principali usi sono dunque i seguenti:

- uso del modello di lessico computazionale sviluppato per la lingua inglese per la costruzione di risorse lessicali per altre lingue. Le principali attività avviate a questo scopo riguardano[20]: lo spagnolo (Subirats e Petruck, 2003)[21]; il giapponese (Ohara et al., 2004)[22]; il tedesco per il quale sono avviati una serie di progetti paralleli i cui principali sono *i)* la costruzione del lessico German FrameNet[23], finalizzato anche alla

[19]Per una descrizione aggiornata e dettagliata delle diverse applicazioni di FrameNet nella comunità di ricerca in materia di Trattamento Automatico del Linguaggio vedi la rassegna di Tonelli (2010, pp. 29–39).

[20]Per una descrizione sempre aggiornata delle attività finalizzate allo sviluppo di lessici computazionali basati sul modello FrameNet vedi la pagina del progetto FrameNet http://framenet.icsi.berkeley.edu/index.php?option=com_content&task=blogcategory&id=94&Itemid=139

[21]http://gemini.uab.es:9080/SFNsite

[22]http://jfn.st.hc.keio.ac.jp/index.html

[23]http://www.laits.utexas.edu/gframenet/

costruzione di un FrameNet bilingue inglese–tedesco (Boas, 2002), e *ii)* il progetto SALSA ("Saarbrücken Lexical Semantics Annotation and Analysis project")[24] (Burchardt et al., 2009), finalizzato alla costruzione di un lessico computazionale a partire da un corpus annotato con informazione semantico–lessicale sulla base dei principi di rappresentazione e organizzazione del significato di FrameNet, realizzato per essere usato in compiti di Trattamento Automatico del Linguaggio; lo svedese (Borin et al., 2009)[25]; il portoghese[26]; l'ebraico moderno (Petruck, 2009); l'italiano, per il quale sono tutt'ora in corso le attività di più università e centri di ricerca coordinate nel progetto IFrame[27].

Sono tutti progetti basati sul riutilizzo del FrameNet costruito per la lingua inglese e allo stesso tempo finalizzati a mettere in luce le specializzazioni (estensioni, ristrutturazioni, ecc...) richieste dalle specificità della nuova lingua;

- uso dei principi teorici e organizzativi di FrameNet per la costruzione di risorse lessicali multilingue, finalizzate alla traduzione automatica (Boas, 2002; Fung e Benfeng, 2004), utilizzando il 'Semantic Frame' come una sorta di interlingua (Boas, 2009).

- in linea con la visione di FrameNet come una 'rete di relazioni', uso delle relazioni 'frame–to–frame' come mezzo di rappresentazione ontologica della conoscenza. Gli esperimenti condotti in questa direzione sfruttano l'organizzazione dell'informazione semantico–lessicale offerta dalle relazioni tra frames per offrire una rappresentazione strutturata del contenuto semantico di un testo a partire dalla sua realizzazione linguistica.

  Tra le applicazioni che seguono questa linea, quelle di maggiore successo sono quelle che hanno raccolto il suggerimento di Fillmore et al. (2004) di focalizzare l'attenzione sulle relazioni che legano un numero ristretto di frames caratterizzanti un determinato dominio.

  Come discusso nel Paragrafo 6.4.1, è questa la strategia seguita da Dolbey (2009, pp. 65–74), finalizzata a mostrare come una completa

[24] http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index
[25] http://spraakbanken.gu.se/eng/swefn
[26] http://www.framenetbr.ufjf.br/
[27] http://sag.art.uniroma2.it/iframe/doku.php?id=start

descrizione dei principali fenomeni biologici contenuti in una collezione di testi di letteratura biomedica sia possibile non solo mettendo in luce l'informazione veicolata dalla rete di frames rappresentativi della semantica di dominio, ma anche trovando un collegamento tra tale rete e le corrispondenti classi ontologiche contenute in ontologie biomediche.

Come effettivamente verificato da Uematsu et al. (2009), l'obiettivo è quello di dimostrare che, proprio grazie ai principi organizzativi di FrameNet, è possibile creare un 'ponte' tra un approccio all'organizzazione della semantica di dominio esclusivamente basato su presupposti teorici di conoscenza di dominio (come quella offerta dalle ontologie di dominio) e uno basato sulla rappresentazione esplicita dell'informazione semantico–lessicale contenuta in una collezione di testi di dominio (offerta dall'annotazione semantica);

- il fatto di considerare FrameNet come 'rete' è il punto di partenza per attività finalizzate a:

  - collegare FrameNet alle classi dell'ontologia formale SUMO ("Suggested Upper Merged Ontology")[28] facendo uso dei tipi di restrizione di selezione semantica espresse dai STs. L'obiettivo in questo senso è quello di creare una risorsa ontologica, in grado cioè di definire in modo formale alcuni principali concetti del mondo, empiricamente basata nello stesso tempo su di una collezione documentale (Scheffczyk et al., 2006a).

    La necessità nasce dalla consapevolezza che i STs in FrameNet sono pochi e organizzati secondo una gerarchia superficiale. Al contrario, la caratterizzazione delle istanze dei FEs in modo ontologicamente fondato apre la strada a nuove applicazioni. Tra queste, quella di maggiore interesse riguarda la possibilità di usare la rete di organizzazione della conoscenza offerta da SUMO per compiti di 'reasoning' automatico a partire da un testo semanticamente annotato secondo i principi di FrameNet (Scheffczyk et al., 2006b). Inoltre, come chiaramente delineato da Scheffczyk et al. (2006a), uno degli obiettivi collegati riguarda la specializzazione di dominio dei STs;

[28] http://www.ontologyportal.org/

– rendere espliciti i rapporti tra eventi espressi in un testo utilizzando le relazioni 'frame–to–frame'. L'obiettivo in questo senso è quello di usare FrameNet per svolgere compiti di gestione automatica dell'informazione contenuta in corpora testuali, quali compiti di co–referenza tra eventi semanticamente collegati (Burchardt et al., 2005; Fillmore et al., 2006). In questo caso l'obiettivo è quello di estendere il paradigma di descrizione del contenuto linguistico e informativo di un singolo periodo all'intero documento, realizzando un vero e proprio compito di comprensione testuale ('Text Understanding') (Fillmore e Baker, 2001);

- uso dei principi di annotazione del testo seguiti nel progetto FrameNet per l'annotazione e l'analisi semantica di collezioni documentali. Tale utilizzo è riconducibile alla doppia finalità perseguita sin dagli esordi del progetto, quella cioè di fornire una risorsa lessicografica ma anche un corpus semanticamente annotato (Lowe et al., 1997).

  L'interesse della comunità di ricerca in materia di Trattamento Automatico del Linguaggio in questo senso è testimoniata dall'uso di FrameNet per compiti di gestione dell'informazione semantica contenuta in corpora testuali, come quelli descritti nei punti precedenti, e per compiti di annotazione semantica automatica (Gildea e Jurafsky, 2002). A partire infatti dal 2004 nell'ambito della campagna di valutazione "Senseval" finalizzata a mettere a confronto sistemi statistici dedicati all'annotazione automatica di ruoli semantici ('Automatic Semantic Role Labeling') nel testo[29], FrameNet è utilizzato come risorsa di riferimento *i)* per lo sviluppo di strumenti di annotazione semantica automatica (Erk e Padó, 2006)[30] e *ii)* per la definizione di metodi di riconoscimento automatico di unità e strutture lessicali che evocano 'Semantic Frames'[31].

  Inoltre, recentemente, in linea con la modalità di annotazione continua del testo, l'attenzione si è spostata sul riconoscimento della relazione anaforica che lega casi di non–istanziazione ('Null Instantiation') di un partecipante ad un evento (frame) con il suo corrispondente istanziato

[29] http://www.senseval.org/senseval3
[30] http://www.coli.uni-saarland.de/projects/salsa/shal/
[31] http://nlp.cs.swarthmore.edu/semeval/tasks/task19/summary.shtml

in un contesto testuale più ampio di quello della singola frase[32]. Viene dunque allargata la prospettiva dell'annotazione semantica, vista non solo "as a sentence internal problem but as a task which should really take the discourse context into account" (Ruppenhofer et al., 2009);

- uso in contesti specialistici per *i)* la costruzione di lessici specialistici e *ii)* l'annotazione semantica di corpora rappresentativi di un determinato linguaggio specialistico, finalizzato alla realizzazione di compiti di gestione della conoscenza di dominio basati su metodi e strumenti di Trattamento Automatico del Linguaggio[33].

6.2 Il confronto con il modello paradigmatico di WordNet

Allo scopo di mettere in luce come i principi di organizzazione del significato adottati da FrameNet siano particolarmente adatti per la rappresentazione del contenuto semantico–lessicale di testi giuridici, in quanto segue tali principi sono messi a confronto con quelli ortogonali sui quali è basato WordNet.

Per chiarezza, è di seguito riportata una breve descrizione dei principi organizzativi e degli elementi principali che compongono WordNet.

6.2.1 I principi e gli elementi organizzativi di WordNet

WordNet[34] è un progetto avviato alla fine degli anni '80 presso l'Università di Princeton da un gruppo di ricerca guidato da George Miller e da Christiane Fellbaum (Fellbaum, 1998), finalizzato allo sviluppo di un lessico computazionale per la lingua inglese. A partire da teorie psicolinguistche sull'organizzazione della memoria lessicale, WordNet si configura come una grande rete semantica, all'interno della quale le parole sono messe in collegamento tra di loro sulla base delle relazioni lessicali e semantiche che le legano.

Alla base vi è l'idea che la memoria semantica di una parola "is not a circle, but a tree (in the sense of tree as a graphical representation)"; ne segue che "the lexical tree can be reconstructed by following trails of superordinate

[32] http://www.coli.uni-saarland.de/projects/semeval2010_FG/

[33] Vedi Paragrafo 6.4.

[34] http://wordnet.princeton.edu/doc

terms: oak @→ tree @→ plant @→ organism, for example, where '@→' is the transitive, asymmetric, semantic relation that can be read 'is a' or 'is a kind of' " (Miller, 1993a, p. 12). In questo modo il lessico è organizzato in un sistema gerarchicamente strutturato di relazioni paradigmatiche tra parole.

WordNet è pertanto basato su principi organizzativi molto simili a quelli di un thesaurus i cui elementi fondamentali sono i seguenti:

- **le parole**: in WordNet è descritto il significato di sostantivi, aggettivi, verbi e avverbi;

- **i synsets**: sono l'unità minima di organizzazione del significato. In WordNet le parole sono organizzate in gruppi di sinonimi ('synonym sets'), definiti tali sulla base del 'principio di sostituzione' in un contesto: "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value" (Miller et al., 1993b, p. 6). Di conseguenza, sostantivi, verbi, aggettivi, avverbi sono organizzati in synsets separati.

  Così, ad esempio, il verbo *to depart* appartiene al synset {go, go away, depart}, composto da altri verbi sinonimi, o il sostantivo *escape* appartiene al synset {escape, flight}.

  Il synset corrisponde di fatto alla nozione tradizionale di 'senso' di una parola. Pertanto, *i)* il synset a cui appartiene la parola x rappresenta il concetto espresso dalla parola x, *ii)* una parola con n–significati appartiene a n–synsets diversi, *iii)* un synset può anche contenere un'unica parola.

  Così, il verbo *to depart*, nel senso di 'partire', appartiene al synset {go, go away, depart}; nel senso di 'divergere', si trova nel synset {deviate, vary, diverge, depart}. Allo stesso modo, il sostantivo *escape*, nel senso di 'atto fisico della fuga', appartiene al synset {escape, flight}; nel senso più specifico di 'evasione', si trova nel synset {escape, evasion, dodging}; o ancora, nel senso molto specifico di 'fuga da una difficoltà o da un pericolo', è l'unico elemento del synset {escape};

- **le glosse e gli esempi**: per chiarezza, ogni synset è accompagnato da una glossa descrittiva del senso espresso da quel determinato insieme di parole e da un esempio trovato sulla base della competenza lessicografica degli sviluppatori di WordNet.

Ad esempio, il synset {go, go away, depart} è accompagnato dalla glossa "move away from a place into another direction" e dai due seguenti periodi di esempio *Go away before I start to cry, The train departs at noon*;

- **le relazioni semantiche** che legano i synsets: sono le relazioni gerarchiche che mettono in collegamento i synsets, definendo in questo modo la struttura 'a rete' di WordNet. Di fatto, il concetto espresso da un synset è pienamente caratterizzato dalla posizione del synset all'interno della rete semantica, dall'insieme cioè delle sue relazioni con altri synsets. Il significato di una parola è dunque rappresentato come un punto discreto in una rete semantica, descritto dalla posizione della parola nella rete di synsets.

  È da notare che a seconda della categoria morfosintattica delle parole contenute in un synset sono preferiti alcuni tipi di relazioni semantiche nella gamma di quelle possibili. In particolare, per i sostantivi sono centrali le relazioni di ipononimia/iperonomia e meronimia/olonimia (Miller, 1993a), mentre i synsets che raggruppano aggettivi sono legati soprattutto dalla relazione di antonimia (Fellbaum et al., 1993a) e i synsets di verbi sono per lo più in relazione di troponimia tra loro (Fellbaum, 1993b).

  Così, ad esempio, il synset di verbi {go, go away, depart} è legato per troponimia al synset {shove off, shove along, blow}, con glossa "leave; informal or rude", che contiene verbi troponimi di *to depart*. Il synset di sostantivi {escape, flight} è legato da una relazione di iperonimia al synset {running away}, con glossa "the act of leaving (without permission) the place you are expected to be", e da una relazione di iponimia al synset {break, breakout, jailbreak, gaolbreak, prisonbreak, prison–breaking}, con glossa "an escape from jail", e a quello {exodus, hegira, hejira}, con glossa "a journey by a large group to escape from a hostile environment"[35].

  È infine importante qui ricordare il fatto che l'intero lessico WordNet è organizzato in un numero limitato di primitive semantiche; ad esempio, i sostantivi sono organizzati in 25 primitive, quali 'food', 'animal',

[35]Nota che ognuno dei due synsets qui riportati come esempi è legato anche da altre relazioni semantiche. Così, ad esempio, il synset {go, go away, depart} è legato per iperonimia al synset {exit, go out, get out, leave}.

'location', 'substance', ecc... (Miller, 1993a). È infatti a partire da questo tipo di organizzazione che sono stati fatti una serie di tentativi per ricondurre le relazioni semantiche esistenti tra i synsets di WordNet a relazioni ontologiche che legano nodi ontologico–concettuali (classi ontologiche) di ontologie formali (Gangemi et al., 2003a; Niles e Pease, 2003).

Un aspetto centrale del modo di organizzare lo spazio semantico–lessicale in WordNet riguarda il trattamento dell'informazione sintattica. I synsets di verbi, oltre all'informazione relativa alle relazioni semantiche con altri synsets verbali, forniscono anche informazioni relative alle proprietà di sottocategorizzazione dei verbi. Per ogni synset verbale viene riportato il tipo di costruzione sintattica ('sentence frame') nella quale i verbi possono ricorrere. Ad esempio, al synset "move away from a place into another direction" a cui appartiene il verbo *to depart* sono associati i seguenti frames sintattici: *Something —-s, Somebody —-s, Something is —-ing PP, Somebody —-s PP.* Ciò implica che tutti i verbi contenuti in questo synset hanno il medesimo comportamento sintagmatico.

Tuttavia, è un tipo di informazione che non ha l'intento di essere esaustiva dal momento che, come ricorda Fellbaum (1993b, p. 55), "WordNet was designed to model lexical memory rather than represent lexical knowledge, so it excludes much of a speaker's knowledge about both semantic and syntactic properties". L'informazione sul comportamento sintagmatico delle parole è infatti associata esclusivamente a synsets di verbi, impedendo pertanto di avere indicazioni sul comportamento sintattico–combinatorio anche di sostantivi, aggettivi e avverbi. È inoltre qui d'interesse mettere in evidenza come si tratti di un'informazione fornita unicamente sulla base di intuizione lessicografica e non a partire dall'analisi di concrete attestazioni in collezioni documentali.

6.2.2 FrameNet vs WordNet: i vantaggi per il dominio giuridico

Una delle domande più frequenti che ricorrono nel forum di domande poste dagli utilizzatori di FrameNet riguarda proprio la relazione tra il progetto FrameNet e WordNet. Come spiegato nelle risposte fornite, l'intenzione originaria degli sviluppatori di FrameNet era quella di creare una risorsa

semantico–lessicale che fosse nei suoi fondamenti teorici e principi organizzativi complementare a WordNet. L'idea era, da un lato, quella di utilizzare le parole contenute nei synsets di WordNet per ampliare la lista di LUs evocatrici di frames; dall'altro, quella di usare l'informazione fornita da FrameNet per espandere gli esempi di WordNet, fondandoli su attestazioni reali, e per aggiungere informazione riguardo al comportamento sintattico–combinatorio delle parole contenute in WordNet.

Sebbene l'obiettivo di trovare un collegamento tra queste due risorse sia al centro delle attività di una serie di gruppi di ricerca[36], sino ad oggi un tale progetto non è ancora stato pienamente realizzato.

Ciò detto, è qui importante focalizzare l'attenzione su un punto cruciale del rapporto tra FrameNet e WordNet, sul fatto cioè che è la codifica di due aspetti ortogonali del significato a rendere strettamente correlati ma anche profondamente diversi i modelli di rappresentazione dell'informazione lessicale adottati nei due progetti. Sebbene infatti per una completa descrizione dello spazio semantico–lessicale sia necessario renderne esplicito il livello di organizzazione sia sintagmatico sia paradigmatico, tuttavia i principi teorici che guidano questi due livelli di rappresentazione del significato sono profondamente diversi.

Partendo da questi presupposti, è qui intenzione mettere in evidenza le principali similarità e divergenze tra i principi organizzativi di WordNet e FrameNet grazie ad una serie di esempi concreti. L'obiettivo è quello di portare l'attenzione su aspetti particolarmente rilevanti in un'ottica di descrizione del significato contenuto in testi giuridici.

Da un lato, le due risorse sono accomunate *i)* dal fatto di fornire un'organizzazione 'a rete' del significato e *ii)* dall'essere utilizzate in compiti di annotazione semantica del testo.

Riguardo al primo aspetto, in entrambe le risorse le relazioni gerarchiche tra synsets in WordNet e tra frames in FrameNet rappresentano il tratto formale caratteristico che consente di considerare i due lessici computazionali come 'ontologie linguistiche'. Sebbene il rapporto tra lessici e ontologie sia un tema controverso e molto dibattuto, come chiarito da Hirst (2003), l'idea di base è che essi siano strettamente correlati e a volte sovrapponibili dal momento che un'**ontologia** è un sistema strutturato di 'oggetti di conoscenza' (concetti) organizzati sulla base delle relazioni che ne costituiscono

[36]Per una rassegna aggiornata vedi Ruppenhofer et al. (2010, p. 86).

l'architettura e un **lessico** è un sistema organizzato di 'oggetti linguistici' tra loro in relazione.

Questo è tanto più vero nel caso di linguaggi espressione di domini specialistici, nei quali la natura settoriale del lessico consente una corrispondenza tra 'oggetti di conoscenza e linguistici' maggiore che nel caso del linguaggio comune. Come discusso da Buitelaar et al. (2009), sono in questo caso particolarmente evidenti, infatti, le potenzialità di utilizzare un'ontologia di dominio per guidare la costruzione di un lessico computazionale specialistico (organizzato cioè sulla base di relazioni semantico–lessicali gerarchiche tra termini chiave per un determinato dominio di conoscenza) e, viceversa, di fondare un'ontologia che organizzi i concetti fondamentali di un dominio su un lessico che ne strutturi le componenti semantico–lessicali rilevanti.

Sia WordNet sia FrameNet, inoltre, sono usati per realizzare compiti di annotazione semantica finalizzati a rendere esplicita l'informazione semantico–lessicale contenuta in collezioni documentali. A partire dai diversi principi di organizzazione del significato su cui sono basati, i due progetti sono tuttavia utilizzati per realizzare compiti di annotazione semantica parzialmente diversi.

FrameNet, fornendo conoscenza relativa all'organizzazione sintagmatica del significato, è infatti usato per lo più (come descritto nel Paragrafo 6.1.3) come risorsa di riferimento per svolgere compiti di annotazione automatica di ruoli semantici ('Automatic Semantic Role Labeling') o di comprensione del testo ('Text Understanding') sulla base della ricostruzione di relazioni di co–referenza semantica tra eventi correlati. WordNet, fornendo conoscenza relativa alle relazioni paradigmatiche tra le parole, è usato soprattutto come 'repertorio di sensi' allo scopo di determinare il significato di un termine in un contesto, disambiguando allo stesso tempo i termini polisemici[37]. Tuttavia, anche nel caso dell'annotazione semantica, WordNet e FrameNet possono essere visti come due risorse di riferimento complementari per una ricca analisi semantica di un testo, come sperimentato da Baker e Fellbaum (2009).

D'altro canto, le due risorse sono ortogonali rispetto ai seguenti aspetti di descrizione e organizzazione del significato:

- **il punto di partenza per la descrizione del significato**: in WordNet i synsets sono definiti sulla base delle competenze lessicografiche

[37]È quanto viene fatto dai sistemi che dal 1997 si confrontano nelle varie edizioni della campagna di valutazione Senseval http://www.senseval.org/past.html

individuali e a partire da dizionari della lingua inglese. FrameNet, invece, essendo sia un lessico computazionale sia un corpus annotato con informazione sintattico–semantica è basato su attestazioni reali in corpora testuali. Di conseguenza, WordNet offre un'ampia descrizione del lessico inglese secondo i metodi lessicografici 'classici', slegata tuttavia dalle concrete modalità d'uso; FrameNet, al contrario, sebbene offra una descrizione del lessico limitato alle occorrenze d'uso della collezione documentale su cui è basato, consente di 'ancorare' la descrizione del significato al testo, rendendo esplicita la relazione tra informazione semantica e comportamento sintattico di un'unità lessicale (Atkins et al., 2003a).

- **l'unità minima di descrizione del significato**: la questione è al centro del ben noto e aperto dibattito su quale modo scegliere per rappresentare il significato lessicale in modo formale e organizzato (Kilgarriff, 1997). WordNet e FrameNet si pongono ai due estremi della discussione.

  Il primo, organizzando le unità lessicali per 'sensi' (synsets), tra loro collegati ma pienamente distinti e differenziati l'uno dall'altro, abbraccia l'idea per cui il significato è descrivibile sotto forma di unità discrete, reciprocamente esclusive, organizzabili in una rete di simboli. Il secondo, organizzando le unità lessicali per schematizzazioni–tipo di situazioni conoscitive (frames), suggerisce come il compito di descrivere il significato consista nel riconoscere gli elementi di conoscenza che contribuiscono a ricostruire un determinato contesto conoscitivo. L'idea fondamentale è che, dal momento che nell'uso linguistico quotidiano i significati sono 'eventi' scomponibili in più componenti semantici e non entità in sè concluse (Hanks, 2000, p. 210), di fatto il significato non esiste al di fuori del contesto d'uso.

  Così, ad esempio, come mostrato nella Figura 6.3, il significato di *to obligate* è univocamente determinato dalla posizione del synset {obligate, bind, hold} (al quale il verbo appartiene) all'interno della rete di relazioni di iperonimia e troponimia nella quale è inserito. In base ai principi composizionali della 'Frame Semantics', il significato di *obligate* è invece determinato a partire dalle diverse componenti semantiche che contribuiscono alla descrizione della situazione conoscitiva a cui il verbo rimanda. Esso è cioè definito sulla base degli elementi caratteriz-

zanti il frame BEING_OBLIGATED che il verbo evoca, quali il 'soggetto tenuto ad adempiere un dovere' (FE Responsible_party), il 'dovere che egli deve adempiere' (FE Duty), il 'luogo nel quale il soggetto deve adempiere il dovere' (FE Place), ecc...;

- **il modo di descrivere il significato di una 'parola' e la definizione stessa di 'parola'**: l'aspetto è strettamente collegato alla questione precedente. In WordNet una 'parola' è intesa come un'unità lessicale il cui significato è descrivibile *i)* nei termini di appartenenza ad un insieme di unità monorematiche sinonime, omogenee rispetto alla categoria morfosintattica (synsets), *ii)* sulla base della posizione del synset di appartenenza all'interno di una rete gerarchicamente organizzata di synsets. Il significato di una parola è dunque dato dalla relativa posizione di un 'senso' in una rete di 'sensi'.

  In FrameNet è diverso innanzitutto il concetto stesso di 'parola', intesa come qualsiasi unità mono e polirematica, dotata di una struttura predicativa e in grado di evocare uno (o più) contesti conoscitivi[38]. Pertanto, il significato di un'unità lessicale predicativa è dato *i)* dall'appartenenza ad un insieme di unità lessicali predicative (LUs) che evocano un frame, dalla sua capacità cioè di rimandare ad una determinata situazione conoscitiva, *ii)* dalla reciproca posizione del frame di appartenza all'interno di una rete di 'situazioni conoscitive'.

  Così, ad esempio, sulla base dei principi di WordNet, il concetto deontico di 'essere legalmente obbligato ad adempiere un dovere' è disperso tra i synsets verbali di cui fa parte il verbo *to obligate* nel suo significato deontico (parte a) e b) della Figura 6.3) e i synsets di aggettivi e sostantivi (rispettivamente parte c) e d) della Figura 6.3).

  Al contrario, i principi di rappresentazione del significato di FrameNet (parte e) della Figura 6.3) consentono di finalizzare il processo di rappresentazione del significato lessicale anche a scopi di rappresentazione della conoscenza. Il fatto che il frame BEING_OBLIGATED, evocato dal verbo *to obligate* (ma non solo), sia inserito in una fitta rete di relazioni 'frame–to–frame' permette di ricostruire l'ampio scenario conoscitivo

[38] Come chiarito da Ruppenhofer et al. (2010, pp. 7–8), in FrameNet per unità **polirematica** si intendono "multiword expressions such as *given name* and hyphenated words like *shut–eye* [...]" e anche "idiomatic phrases such as *middle of nowhere* and *give the slip (to)*".

di cui il frame fa parte, il quale a sua volta offre una descrizione del concetto di 'obbligo';

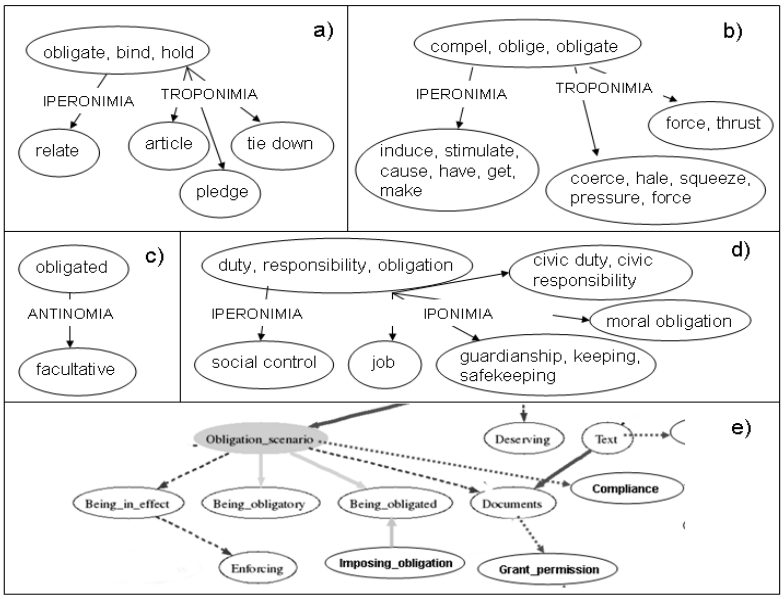


Figura 6.3: La rappresentazione del significato di *to obligate*, di *obligated* e *obligation* in WordNet, parti a), b), c) e d), e in FrameNet, parte e).

- **l'analisi delle diverse categorie morfosintattiche**: in WordNet ogni categoria morfosintattica è considerata separatamente. Come messo in evidenza dai suoi stessi creatori, è in seguito a questa organizzazione del lessico per synsets omogenei rispetto alla categoria morfosintattica che WordNet preclude un approccio completo al comportamento combinatorio di un'unità lessicale (Fellbaum, 1998).

  Di conseguenza, come discusso precedentemente, ad esempio, la descrizione del concetto di 'obbligo' è suddivisa nella distinta rappresentazione del significato dei verbi (parte del synset {obligate, bind, hold}), sostantivi (parte del synset {duty, responsibility, obligation}) e aggettivi (parte del synset {obligated}) che lo realizzano lessicalmente.

  In FrameNet, invece, l'organizzazione del significato per frames consente di considerare unitamente una lista di unità lessicali diverse per ca-

tegoria morfosintattica, ma omogenee rispetto ai componenti semantici che contribuiscono a definire un determinato contesto conoscitivo.

Questo approccio permette così di raggruppare anche unità polirematiche e costruzioni supporto che condividono lo stesso contesto d'uso. Il frame BEING_OBLIGATED è infatti contemporaneamente evocato da verbi (*gotta, hafta, have_to*, ecc...), sostantivi (*assignment, responsibility, contract*, ecc...), aggettivi (*bound, obligated*, ecc...), ma anche da costruzioni a verbo supporto (come ad esempio *[have]responsibility, [claim]responsibility, [entrust]task*, dove l'informazione semantica è veicolata dai sostantivi *responsibility* e *task*) o a preposizione supporto (come ad esempio *[on]responsibility*, dove l'informazione semantica è veicolata dal sostantivo *responsibility*);

- **il modo in cui vengono trattate sinonimia e polisemia**: in WordNet la relazione di sinonimia è la relazione costitutiva del synset. Pertanto, *i)* due parole sono sinonime se appartengono allo stesso synset (cioè se possono essere liberamente sostituite in una frase senza alterarne l'accettabilità), *ii)* due parole sono polisemiche se appartengono a più di un synset.

  Ad esempio, il verbo *to obligate* è sinonimo dei verbi *to compel, to oblige* contenuti nello stesso synset con glossa "force somebody to do something" ed è polisemico dal momento che appartiene anche al synset con glossa "commit in order to fulfill an obligation" e a quello con glossa "bind by an obligation; cause to be indebted".

  FrameNet, al contrario, non consente di rendere conto della relazione di sinonimia in modo 'classico'. In FrameNet infatti due unità lessicali sono semanticamente simili (sinonime) se evocano la stessa situazione d'uso (frame).

  Quindi, ad esempio, il verbo *to adhere*, il sostantivo *adherence*, l'unità polirematica *in accordance*, ecc... ma anche 'parole' che in WordNet sarebbero antonime delle precedenti come il verbo *to violate*, il sostantivo *contravention*, ecc... sono incluse nella lista di unità lessicali evocatrici del frame COMPLIANCE. Il verbo *to adhere* è polisemico perché appartiene alla lista di unità lessicali evocatrici dei frames COMPLIANCE, ATTACHING, BEING_ATTACHED.

  In questo modo, in base ai principi teorici e organizzativi seguiti, FrameNet consente di rendere esplicita la relazione di **parafrasi** che le-

ga due o più unità lessicali predicative evocatrici di uno stesso frame (Ruppenhofer et al., 2010)[39], più che la relazione di sinonimia che le lega. Questo è esplicitamente indirizzato a "many of the other goals of semantic NLP, including Question Answering, Summarization, and Translation" (Ruppenhofer et al., 2010, p. 85). Compiti di questo tipo sarebbero infatti facilitati da una fase di annotazione che permettesse di rendere esplicita la relazione tra tutte le 'parole' che in un testo rimandano allo stesso contesto conoscitivo (frame).

Un esempio particolarmente significato ai fini di questo lavoro è discusso da Fillmore e Baker (2010, pp. 335-336). Si tratta del frame COMPLIANCE del quale sono considerate parafrasi ugualmente evocatrici costruzioni diverse che appartengono a più di una categoria morfosintattica (es. *This* ***conforms*** *to the regulation/is* ***in conformity*** *with the regulation/is* ***compliant*** *with the regulation*), casi di antonimia (es. *This* ***conforms*** *to the regulation/is* ***in violation*** *of the regulation/is* ***not in compliance*** *with the regulation*), ecc...[40]

Di conseguenza, anche la relazione di antonimia non viene trattata in modo tradizionale. L'informazione relativa alla presenza di antonimi evocatrici di un medesimo frame è resa esplicita specificando all'interno delle LUs se si tratta di volta in volta di una 'Positive' o 'Negative' LU (Ruppenhofer et al., 2010, p. 84).

Come discusso nel Capitolo 7, questo trattamento dell'antonimia non sempre è soddisfacente nel caso della rappresentazione di concetti deontici[41];

- **il trattamento dell'informazione sintattica**: in base ai principi teorici e organizzativi di WordNet le proprietà sintattico combinatorie delle parole non sono oggetto di rappresentazione. Sono contenute nel database lessicale unicamente informazioni generali relative ai possibili frames di sottocategorizzazione verbale di synsets di verbi.

[39] "One of the basic insights behind FrameNet is that grouping words according to the scenes that they evoke, regardless of whether they are synonyms, antonyms, or some other relation to each other, groups words that are useful for paraphrasing. In particular, since FrameNet lists words together despite part–of–speech differences (unlike WordNet), paraphrases involving an interchange of noun, verb, adjective, or preposition are (in principle) discoverable with the FrameNet data." (Ruppenhofer et al., 2010, p. 85).

[40] Sono riportate in grassetto le parole o costruzioni evocatrici un frame.

[41] Vedi Paragrafo 7.6.3.1.

Così, ad esempio, come riportato nella Figura 6.4, al synset {oblige, bind, hold, obligate} sono associati tre diversi 'sentence frames' condivisi da tutti i verbi sinonimi parte del synset.

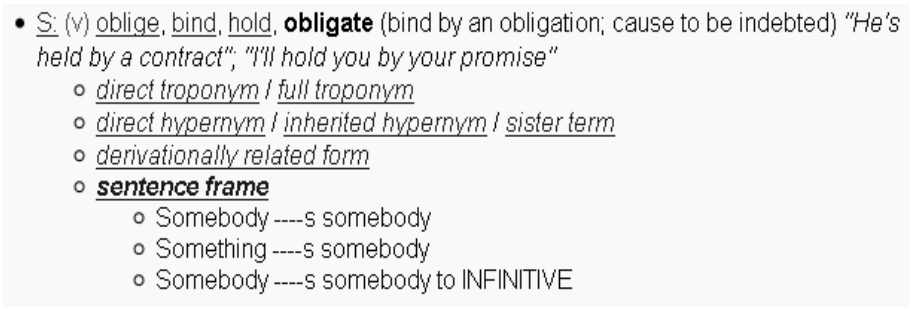


Figura 6.4: I tre diversi 'sentence frames' associati ai verbi parte del synset {oblige, bind, hold, obligate}.

Al contrario, in FrameNet l'informazione relativa ai vincoli sintattico–combinatori tra le parole è parte integrante nella descrizione di tutte le LUs a prescindere dalla loro categoria morfosintattica. A differenza di WordNet infatti *i)* ogni singola LUs ha una serie di strutture sintattiche associate e *ii)* le strutture presenti nel database di FrameNet sono le diverse realizzazioni sintattiche rintracciate nel British National Corpus. Come mostrato nella Figura 6.5, per ogni LU sono raccolte tutte le diverse realizzazioni sintattico–funzionali dei FEs annotati.

Questa caratteristica rende FrameNet un modello di rappresentazione del significato particolarmente espressivo ai fini di uno studio che mira a fornire una descrizione dei rapporti tra struttura sintattico–grammaticale di periodi giuridici e il modo in cui vi è organizzato il contenuto semantico–informativo.

6.3 Il confronto con altri progetti di rappresentazione sintagmatica del significato

Allo scopo di completare la descrizione dei motivi che hanno guidato la scelta di FrameNet come modello di riferimento per la rappresentazione del contenuto semantico di testi giuridici, sono qui di seguito passati in rassegna gli

Frame Elements and Their Syntactic Realizations

The Frame Elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Condition	(3)	PP[under].Dep (3)
Duty	(5)	DEN.-- (3) N.Dep (2)
Responsible_party	(6)	Poss.Gen (5) NP.Ext (1)

Figura 6.5: La realizzazione sintattica dei FEs per la LU *obligation* evocatrice del frame BEING_OBLIGATED.

altri principali progetti oggi esistenti basati su principi di rappresentazione e organizzazione a livello sintagmatico del significato lessicale.

Per chiarezza, è stata fatta una distinzione tra i progetti basati sull'annotazione semantica di corpora e VerbNet, l'unico progetto di questo tipo non fondato su evidenza testuale.

6.3.1 Progetti basati sull'annotazione semantica di corpora

Il progetto PropBank[42]

Attivo presso l'Università del Colorado (Boulder), il progetto è finalizzato alla costruzione di una collezione di proposizioni a partire dalla struttura predicato–argomenti di **verbi** della lingua inglese (Palmer et al., 2005). A questo scopo, è stata utilizzata come risorsa di riferimento la Penn TreeBank–II [43], il repertorio di periodi del Brown Corpus e del Wall Street Journal linguisticamente annotati fino al livello sintattico.

Il progetto è focalizzato, in particolare, sull'annotazione dei ruoli semantici associati agli argomenti sintattici parte dei frames di sottocategorizzazione dei verbi presenti nella Penn TreeBank–II. Ad esempio, come mostrato in quanto segue, i due sensi del verbo *to execute* ('uccidere' e 'promulgare')

[42]http://verbs.colorado.edu/∼mpalmer/projects/ace.html
[43]http://www.cis.upenn.edu/∼treebank/

i) sono descritti grazie ad una serie di argomenti (Arg) numerati in modo progressivo (da 0 a 5), ai quali viene poi associato un nome mnemonico sulla base del senso corrispondente, e *ii)* il loro uso è esemplificato da una frase della Penn TreeBank–II i cui componenti sintattici sono stati annotati con i corrispondenti argomenti semantici:

execute.v(0.1) : kill

Arguments :

Arg0 : killer

Arg1 : corpse

Arg2 : instrument

Example : *John **executed** the criminal with his umbrella.*

Arg0 : John

Arg1 : the criminal

Arg2 : with his umbrella

execute.v(0.2) : enact

Arguments :

Arg0 : agent, doer

Arg1 : thing done

Arg2 : benefactive, done for or about

Arg3 : instrumental

Example : *Mr. Allen's Pittsburgh firm, Advanced Investment Management Inc., **executes** program trades for institutions.*

Arg0 : Mr. Allen's Pittsburgh firm, Advanced Investment Management Inc.

Arg1 : program trades

Arg2 : for institutions

Il progetto NomBank[44]

Collegato al progetto PropBank e basato sulla Penn TreeBank–II, il progetto è attivo presso la New York University ed è finalizzato alla costruzione di una collezione di proposizioni a partire dalla struttura predicato–argomenti

[44]http://nlp.cs.nyu.edu/meyers/NomBank.html

di **sostantivi deverbali**, sostantivi cioè per i quali esiste un corrispondente verbo nella PropBank (Meyers et al., 2004).

Realizzato sulla base di NOMLEX ("NOMinalization Lexicon")[45] (Macleod et al., 1998), il progetto ha l'obiettivo di rendere esplicita la relazione tra gli elementi (gli argomenti) del frame di sottocategorizzazione di ogni verbo (nei suoi diversi sensi) e i corrispondenti modificatori parte della struttura argomentale del sostantivo deverbale derivato[46]. Ad esempio, per ogni struttura predicato–argomenti associata ad ognuno dei due sensi del verbo *to execute* nella PropBank, nell'ambito del progetto NomBank è stato trovato un collegamento con ogni singolo argomento della struttura argomentale del sostantivo *execution*. Come mostrato in quanto segue, ad esempio, l''Arg1, corpse' del verbo corrisponde all'Arg1, corpse' del sostantivo, sintatticamente realizzato come sintagma preposizionale modificatore di *execution*:

execution.n(0.1) (source="verb–execute.01")

Arguments :

Arg0 : killer

Arg1 : corpse

Arg2 : instrument

Example : *the* ***execution*** *of mass–murderer Ted Bundy – who eventually was executed*

Arg1 : of mass–murderer Ted Bundy – who eventually was executed

executio.n(0.2) (source="verb–execute.02")

Arguments :

Arg0 : agent, doer

Arg1 : thing done

Arg2 : benefactive, done for or about

Arg3 : instrumental

Example : *the president's* ***execution*** *of the law*

Arg0 : the president's

Arg1 : of the law

[45] http://nlp.cs.nyu.edu/nomlex/index.html

[46] Come nella PropBank, anche nella NomBank ogni struttura predicato–argomenti associata ad un senso di un sostantivo è accompagnata da un esempio tratto dalla Penn TreeBank–II.

Il progetto "Corpus Pattern Analysis"[47]

Attivo presso la Brandeis University, è promosso da Patrick Hanks e dedicato alla costruzione di una nuova generazione di vocabolari basati sulla raccolta delle strutture sintagmatiche prototipiche nelle quali ricorrono tipicamente le parole in corpora testuali.

In particolare, il progetto è finalizzato alla costruzione del "Pattern Dictionary of English Verbs" (Hanks, 2008), ad oggi in via di compilazione, che comprende per ogni lemma verbale una lista di tutti i più frequenti contesti d'uso nel British National Corpus, associati a un determinato senso del lemma. Nel dizionario il contesto di ogni verbo è descritto da una serie di 'patterns' definiti sulla base dei ruoli semantici (specifici per ogni contesto) e dei tipi semantici associati alle parole parte del contesto. Al verbo *to grasp*, ad esempio, nel senso di 'afferrare' sono associati tre diversi 'patterns', con le rispettive frequenze d'occorrenza nel corpus di partenza:

(a) [[Person=Animate]] **grasp** [[PhysObj]] (14%)

(b) [[Person 1=Animate]] **grasp** [[Person 2=Animate]] (by[[BodyPart|Clothing]]) (13%)

(c) [[Person=Animate]] **grasp** [[NO OBJ]] (at|for) [[PhysObj]] (2%)

La finalità di questo progetto è quella di creare una risorsa lessicografica complementare a FrameNet. Espressamente fondata sull'idea che il significato di una parola sia pienamente determinato dal contesto nel quale essa ricorre, la "Corpus Pattern Analysis" considera tuttavia come un unità primaria di analisi il **senso** di ogni singolo verbo invece del 'Semantic Frame'. Ciò permette di superare uno dei limiti maggiori di FrameNet, che procedendo 'frame per frame' e senza una preliminare analisi delle occorrenze d'uso nell'intero corpus di riferimento (senza una 'corpus pattern analysis') corre il rischio di mettere le diverse istanze di un determinato frame tutte sullo stesso piano, senza fare cioè distinzioni rispetto alla significatività (prototipicità) di un'istanza rispetto alla totalità delle istanze. Al contrario, il progetto lessicografico di Hanks, fondato sulla 'Theory of Norms and Exploitations', per ogni senso di ogni singolo lemma "discovers the normal patterns, sets aside exploitation and other oddities, and attaches a meaning [...] to each normal pattern" (Hanks e Pustejovsky, 2005).

[47] http://nlp.fi.muni.cz/projekty/cpa/#hanks_2004

6.3.2 VerbNet

È un lessico verbale organizzato in classi di verbi semanticamente omogenei che condividono lo stesso comportamento sintattico (Kipper-Schuler, 2005)[48]. Modellato estendendo la classificazione realizzata da Levin (1993), VerbNet è dunque fondato sull'idea che le proprietà sintattico–combinatorie condivise da gruppi di verbi siano diretta espressione della loro semantica. Partendo da questo presupposto, il principio di classificazione si basa sulla capacità di un verbo di ricorrere in più di una struttura sintattica mantenendo lo stesso significato (fenomeno a cui Levin fa riferimento con il nome di 'diathesis alternations').

Come mostra la Figura 6.6, che ne riporta un esempio, ogni classe di VerbNet si presenta come un insieme di verbi al quale è associata una lista di ruoli tematici generali (del tipo 'agente', 'paziente', ecc...) svolti dagli elementi parte della struttura argomentale condivisa da tutti i verbi membri della classe[49]. Il verbo *to execute*, ad esempio, nel senso di 'uccidere', fa parte della classe 'Murder' insieme ai verbi *to assassinate, to eliminate*, ecc..., con i quali condivide gli stessi ruoli 'Agent', 'Patient' e 'Instrument' e le stesse strutture sintattiche (Frames) nelle quali può occorrere, insieme ad un esempio fittizio.

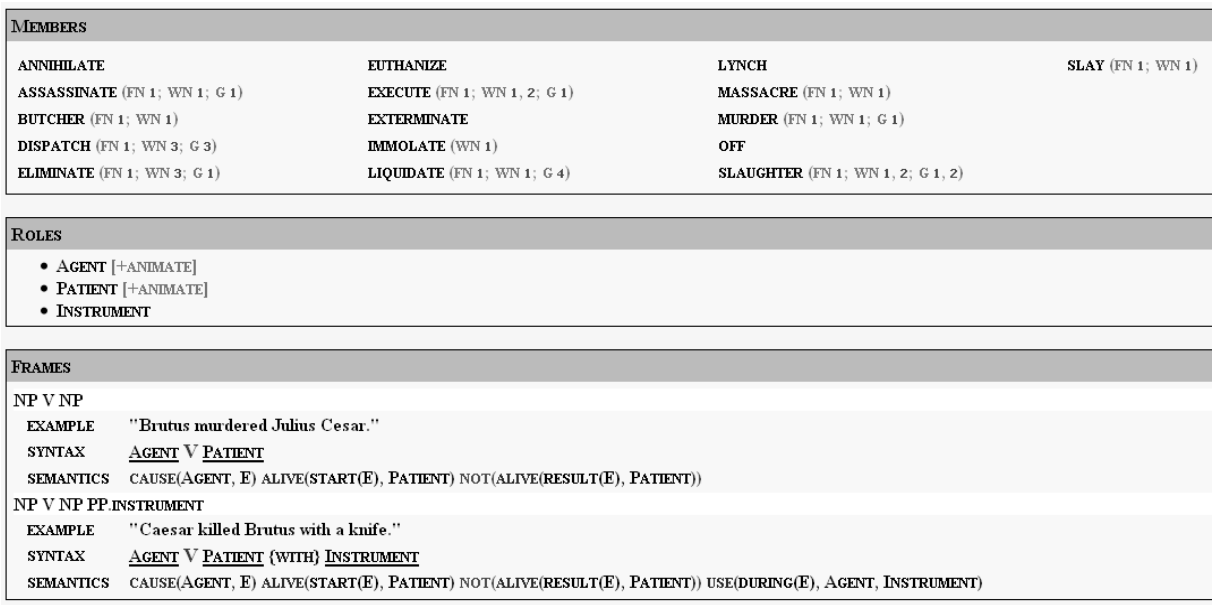
Come si può vedere nella Figura 6.6, inoltre, per ogni verbo della classe 'Murder' è trovata una corrispondenza con lo stesso verbo in WordNet e FrameNet. Nel primo caso, ogni verbo è collegato al synset di WordNet corrispondente al senso che esso ha nella classe di VerbNet. Nel secondo caso, la connessione con FrameNet è realizzata a due livelli, attraverso il collegamento (quando esistente) *i)* tra ogni membro di una classe e un 'Semantic Frame' di FrameNet e *ii)* tra i ruoli tematici di VerbNet e i FEs di FrameNet[50].

Dal momento che VerbNet non è stato costruito a partire da annotazioni testuali, Kipper-Schuler (2005) ha ritenuto opportuno verificare la copertura delle strutture sintattiche classificate sulla base dell'intuizione lessicografica su un corpus di reali attestazioni d'uso. A questo scopo, la loro effettiva

[48] http://verbs.colorado.edu/∼mpalmer/projects/verbnet.html

[49] Grazie alle restrizioni di selezione semantico–ontologica associate ai ruoli tematici, VerbNet è messo in collegamento con i nodi concettuali 'alti' di EuroWordNet (http://www.illc.uva.nl/EuroWordNet/), estensione multilingue di WordNet (Vossen, 1998). Come mostrato nella Figura 6.6, ad esempio, il ruolo 'agent' della classe 'Murder' deve appartenere ad un attore della classe ontologica 'animate'.

[50] Per la descrizione dettagliata dei diversi tipi di collegamento vedi Kipper-Schuler (2005).



MEMBERS

ANNIHILATE	EUTHANIZE	LYNCH	SLAY (FN 1; WN 1)
ASSASSINATE (FN 1; WN 1; G 1)	EXECUTE (FN 1; WN 1, 2; G 1)	MASSACRE (FN 1; WN 1)	
BUTCHER (FN 1; WN 1)	EXTERMINATE	MURDER (FN 1; WN 1; G 1)	
DISPATCH (FN 1; WN 3; G 3)	IMMOLATE (WN 1)	OFF	
ELIMINATE (FN 1; WN 3; G 1)	LIQUIDATE (FN 1; WN 1; G 4)	SLAUGHTER (FN 1; WN 1, 2; G 1, 2)	

ROLES

- AGENT [+ANIMATE]
- PATIENT [+ANIMATE]
- INSTRUMENT

FRAMES

NP V NP

EXAMPLE "Brutus murdered Julius Cesar."

SYNTAX AGENT V PATIENT

SEMANTICS CAUSE(AGENT, E) ALIVE(START(E), PATIENT) NOT(ALIVE(RESULT(E), PATIENT))

NP V NP PP.INSTRUMENT

EXAMPLE "Caesar killed Brutus with a knife."

SYNTAX AGENT V PATIENT {WITH} INSTRUMENT

SEMANTICS CAUSE(AGENT, E) ALIVE(START(E), PATIENT) NOT(ALIVE(RESULT(E), PATIENT)) USE(DURING(E), AGENT, INSTRUMENT)

Figura 6.6: La classe di verbi 'Murder' in VerbNet.

occorrenza è stata ricercata nella PropBank, dove nell'84% dei casi è stata rintracciata una corrispondente struttura predicato–argomenti.

6.3.3 Gli aspetti complementari

Con l'intento di mostrare come i diversi progetti, sebbene basati su principi di organizzazione e rappresentazione del significato diversi, siano messi in collegamento, è qui di seguito brevemente descritto il progetto avviato a questo preciso scopo: il progetto SemLink[51].

L'intento che ha mosso la realizzazione di questa iniziativa è riconducibile alla volontà di armonizzare gli aspetti complementari delle diverse risorse semantico–lessicali esistenti mettendone in luce i singoli vantaggi e superandone in questo modo le limitazioni. Il progetto è pertanto finalizzato a costruire un'ampia base di conoscenza semantico–lessicale che contenga unitamente tutti i diversi tipi di informazioni, sino ad oggi raccolte, sul lessico della lingua inglese.

[51] http://verbs.colorado.edu/semlink/

SemLink, mettendo in collegamento le informazioni contenute nelle diverse risorse per quanto riguarda i **verbi**, ha dato origine ad un database lessicale di verbi: il "Unified Verb Index"[52], che permette di avere informazioni circa:

- la relazione esistente tra i vari sensi di un singolo verbo e il corrispondente comportamento sintattico-combinatorio, in termini di struttura predicato–argomenti (a partire dall'informazione contenuta nella PropBank);
- la semantica condivisa da verbi che hanno lo stesso comportamento sintattico (sulla base della classificazione proposta in VerbNet);
- gli elementi semanticamente rilevanti che consentono ad un gruppo di verbi di descrivere uno scenario conoscitivo condiviso (a partire dai FEs dei 'Semantic frames' di FrameNet);
- quali verbi tra loro sinonimi definiscono un determinato concetto (sulla base della lista di verbi parte di un dato synset di WordNet).

È da notare che questo collegamento è stato realizzato a livello sia del lemma verbale sia del singolo senso di ogni verbo. Così, ad esempio, la semantica del verbo *to execute*, nel senso di 'put to death', è descritta nei termini *i)* del predicato *execute.01* della PropBank, con la sua lista di argomenti numerati; *ii)* di appartenenza alla classe 'Murder' di VerbNet; *iii)* di appartenenza al 'Semantic Frame' EXECUTION; *iv)* di appartenenza al synset {execute}, con glossa "kill as a means of socially sanctioned punishment", e al synset {execute}, con glossa "murder in a planned fashion"[53].

Come ricordato in (Shi e Mihalcea, 2005; Giuglea e Moschitti, 2006), l'obiettivo applicativo di tale iniziativa è quello di creare un'ampia base di conoscenza che, codificando aspetti complementari dell'informazione semantico–lessicale, possa contribuire allo sviluppo di sistemi computazionali avanzati, in grado cioè di svolgere compiti di analisi semantica automatica di testi ('semantic parsing') e, in particolare, di annotazione automatica di ruoli semantici ('Automatic Semantic Role Labeling') con risultati migliori dei sistemi costruiti sulla base di una risorsa lessicale che codifica un solo aspetto della conoscenza semantico–lessicale.

[52]http://verbs.colorado.edu/verb–index/

[53]La stessa raccolta di informazioni è disponibile per il verbo nei suoi altri due sensi previsti: quello di 'do, put into effect, or carry out terms or intent' e quello di 'sign a legal document before witnesses '.

6.3.4 FrameNet vs gli altri progetti: i vantaggi per il dominio giuridico

Fatti salvi gli aspetti complementari dei progetti lessicografici presentati, è qui d'interesse evidenziarne piuttosto le differenze.

È prima di tutto importante ricordare che la definizione dei modelli di rappresentazione del significato adottati in tali progetti è guidata da presupposti linguistico–teorici profondamente diversi che hanno ricadute applicative diverse. Come ricordato da Ellsworth et al. (2004), è il caso, ad esempio, dell'utilizzo di questi modelli in fase di annotazione semantica del testo. Nel momento in cui modelli diversi di rappresentazione del significato vengono applicati alla codifica dell'informazione semantico–lessicale contenuta in un testo essi si realizzano in schemi diversi di annotazione semantica che influenzano il tipo di ricerche condotte sul testo diversamente annotato.

È questo l'orizzonte di questo lavoro. La disamina delle differenze mira infatti a mettere in luce i vantaggi che ha il modello sviluppato nell'ambito del progetto FrameNet se usato nell'annotazione semantica di testi giuridici.

In un'ottica dunque di annotazione semantica del testo, le maggiori variazioni tra i modelli descritti riguardano i seguenti aspetti:

- **la categoria morfosintattica delle unità lessicali di cui rappresentare il significato**: mentre nei progetti PropBank, VerbNet, NomBank e in quello legato alla "Corpus Pattern Analysis" è previsto che la descrizione del significato sia relativa ad una singola categoria morfosintattica, in base al modello FrameNet unità lessicali predicative che appartengono a categorie morfosintattiche diverse possono evocare lo stesso frame (e dunque condividere lo stesso significato). Questo permette un maggior livello di astrazione in fase di annotazione semantica.

  Questo permette di rendere esplicito, per esempio, quando in un atto normativo si sta facendo riferimento agli obblighi che un gestore di impianti di combustione è tenuto ad adempiere grazie all'annotazione di tutti i sostantivi, verbi, aggettivi, avverbi che in un testo 'evocano' una tale situazione di obbligo;

- **l'organizzazione dell'unità minima di rappresentazione del significato**: rispetto a ciò, FrameNet si differenzia dagli altri progetti per l'organizzazione 'a rete' dei 'Semantic Frames'. Assente nella PropBank e NomBank e non prevista dalla "Corpus Pattern Analysis",

questo tipo di organizzazione è presente in VerbNet dove è tuttavia finalizzata a scopi diversi.

Differenziandosi dunque dagli altri progetti sotto questo aspetto, la modalità di organizzazione adottata in FrameNet permette, in fase di annotazione semantica, di rendere esplicito come in un testo una determinata LU rimandi alla rete di altre situazioni–tipo nella quale una determinata situazione conoscitiva evocata è inserita.

Così, ad esempio, l'annotazione in un atto normativo delle proprietà combinatorie (sintattiche e semantiche) del verbo *to obligate*, evocatore del frame IMPOSING_OBLIGATION, permetterà non solo di rendere esplicito il fatto che nel periodo ci si sta riferendo ad una situazione nella quale 'qualcuno o qualche principio regolativo impone un obbligo a qualcuno', ma anche che (sulla base delle relazioni 'frame–to–frame' previste)[54] una tale situazione è inserita in un più ampio scenario di obbligo che prevede altre situazioni–tipo correlate, quali il fatto che 'qualcuno sia obbligato ad adempiere ad un obbligo', il fatto che 'uno stato di cose sia obbligatorio', ecc...;

- **la tipologia di ruoli semantici**: la questione è strettamente legata agli obiettivi dei diversi progetti. Finalizzati a creare una risorsa tale da poter essere utilizzata con successo in compiti di annotazione automatica di ruoli semantici ('Semantic Role Labeling'), i progetti PropBank e NomBank prevedono una lista di argomenti numerati, assumendo così una posizione, che, sebbene controversa, risulta neutrale nell'aperto dibattito sui diversi approcci teorici al tema del rapporto tra ruoli semantici e argomenti sintattici (Levin e Hovav, 1996).

  In linea con l'obiettivo di catturare generalizzazioni di comportamento sintattico che siano un riflesso della semantica sottostante, in VerbNet è utilizzata una serie di ruoli tematici il più generale possibile, ruoli che siano cioè condivisi da tutti i membri di ogni classe (Kipper-Schuler, 2005, pp. 30–35).

  La scelta dei ruoli adottati dalla "Corpus Pattern Analysis" e in FrameNet è guidata da criteri simili tra loro, legati alla semantica specifica dell'unità minima di rappresentazione del significato. Ciò implica che nel primo caso i ruoli semantici sono definiti sulla base dello specifico

[54] Vedi la parte e) della Figura 6.3.

contesto d'uso ('pattern') nel quale ricorre un singolo verbo e sono dunque diversi per ogni senso del verbo. Nel secondo caso, i ruoli semantici (FEs) sono definiti sulla base del 'Semantic Frame' di cui fanno parte e sono pertanto condivisi da tutte le LUs che evocano il frame. Nel primo caso, dunque, i ruoli semantici sono definiti sulla base del **significato**, nel secondo caso sulla base della **situazione–tipo**.

In fase di annotazione semantica, l'adozione in FrameNet di etichette di descrizione dei FEs (per lo più) specifiche per ogni frame (almeno nel caso dei 'Core' FEs) contribuisce alla piena caratterizzazione della semantica di una situazione. Così, ad esempio nel seguente periodo annotato sulla base del modello FrameNet, è reso esplicito che *i punti vendita* svolgono il ruolo tematico generico di 'pazienti', ma nella situazione–tipo specifica (il frame BEING_OBLIGATED), evocata dal participiato passato *obbligato*, giocano il ruolo specifico di 'soggetti tenuti ad adempiere il dovere' (FE Responsible_party):

- [BEING_OBLIGATED] [*Qualora, in attuazione delle disposizioni del comma 2, siano avviate al consumo in rete miscele combustibile diesel–biodiesel con contenuto in biodiesel in misura superiore al 5 per cento* $_{Condition}$], [*i punti vendita nei quali tali miscele sono distribuite* $_{Responsible\_party}$] *sono* ***obbligati*** [*ad esporre idonee etichette di descrizione del prodotto, unitamente all'elenco dei veicoli omologati per l'uso dei predetti biocarburanti* $_{Duty}$].

- **l'interfaccia sintassi/semantica**: in questo caso la questione riguarda principalmente PropBank (e NomBank) e FrameNet, dal momento che nell'approccio della "Corpus Pattern Analysis" all'annotazione semantica non è prevista una fase preliminare di annotazione sintattica del testo.

  Le differenze sono riconducibili alle finalità, profondamente diverse, dei progetti. Entrambi i progetti (PropBank e NomBank) finalizzati all'annotazione del contenuto proposizionale della Penn TreeBank–II aggiungono un livello di annotazione semantica al precedente livello di annotazione sintattica, completando in questo modo il processo stratificato di annotazione linguistica del testo. Di conseguenza, ogni argomento parte della struttura predicato–argomenti annotata è associato ad un nodo dell'albero sintattico sottostante (della Penn TreeBank–II).

Questo non avviene invece in FrameNet, dove le istanze testuali dei FEs non sono limitate in modo restrittivo all'annotazione sintattica del testo. Di fatto, in FrameNet il collegamento sintassi/semantica è reso esplicito a livello 'locale', ma non 'globale'. L'informazione sintattica è cioè relativa ai singoli costituenti sintattici in cui si istanziano i FEs e alle funzioni grammaticali che legano i FEs alla LU evocatrice, a prescindere dalla struttura sintattica dell'intera frase annotata.

Come osservato da Dolbey (2009, p. 23), l'assenza di informazione sulla struttura sintattica globale della frase può creare alcune difficoltà nello svolgimento di compiti computazionali, quali ad esempio l'annotazione automatica di ruoli semantici ('Semantic Role Labeling'). È questo il motivo per cui in questo lavoro si è deciso di mettere a punto una metodologia di annotazione semantica dei testi giuridici che, pur basata sul modello FrameNet, permetta di rendere esplicita la relazione tra realizzazione sintattica globale del periodo e contenuto proposizionale[55];

- **i criteri di scelta dei periodi da annotare**: anche in questo caso, la questione riguarda soprattutto i modelli espressamente finalizzati all'annotazione semantica del testo. Il progetto PropBank prevede che siano annotate a livello semantico tutte le strutture sintattiche nelle quali ricorrono **tutti** i verbi della Penn TreeBank–II nei diversi sensi. Pertanto la PropBank offre una collezione di periodi annotati, esaurientemente rappresentativa delle possibili combinazioni sintassi/semantica nel corpus di partenza, con l'esplicito obiettivo "for the first time to determine the frequency of syntactic variantions in practice" (Palmer et al., 2005). Sulla scia di questo progetto, nella NomBank sono annotati **tutti** i periodi della Penn TreeBank–II che contengono istanze di sostantivi deverbali per i quali ci sia una corrispondente istanza verbale annotata.

  In base al modello della "Corpus Pattern Analysis", il cui obiettivo primario è quello di "account for all normal meanings of each word" (Hanks e Pustejovsky, 2005), la scelta dei periodi da usare come concreti esempi d'uso di 'patterns' è determinata dalla significatività della loro frequenza nel British National Corpus, dal loro cioè essere esemplari prototipici di una norma d'uso.

[55] Vedi Paragrafo 7.2.

In FrameNet, invece, i periodi annotati del British National Corpus non sono scelti perché rivelatori di proprietà combinatorie (sintattiche e semantiche) prototipiche di una LU evocatrice di un determinato 'Semantic Frame'. Ciò è in linea con le finalità che guidano il processo di annotazione in FrameNet, espressamente finalizzato a non fornire indicazioni sulla frequenza d'uso delle informazioni sintattiche e semantiche raccolte nel database lessicale (Ruppenhofer et al., 2010)[56]. È questa anche la finalità del caso di studio presentato in questo lavoro, indirizzato a suggerire una innovativa modalità di annotazione semantica di testi giuridici piuttosto che i risultati di un esaustivo processo di annotazione.

6.4 Utilizzo di modelli di rappresentazione del significato in domini specialistici

Più che nel caso della lingua comune, nella rappresentazione dello spazio semantico–lessicale contenuto in una collezione di testi di dominio l'adozione di modelli di organizzazione del significato è di estrema utilità. Essa permette infatti di rendere esplicito il modo idiosincratico in cui il lessico è espressione della semantica di un dominio di conoscenza. Tali modelli consentono, ad esempio, di portare l'attenzione su unità lessicali espressione di concetti o situazioni specifiche di un dominio che hanno comportamenti sintattico–semantici diversi dalla lingua comune o che non figurano affatto nel lessico comune.

Per questo motivo sono state avviate a livello internazionale una serie di iniziative finalizzate alla costruzione di risorse semantico–lessicali di dominio. Tali risorse si configurano per lo più come estensioni e specializzazioni dei modelli di rappresentazione già esistenti per la lingua comune. L'obiettivo condiviso è quello di mostrare come un modello formale in grado di rendere esplicito il rapporto tra uso della lingua comune e organizzazione della semantica di dominio possa essere utile sia per uno studio delle caratteristiche di un linguaggio specialistico sia per lo sviluppo di sistemi di elaborazione automatica di testi rappresentativi di tale linguaggio. La questione è strettamente

[56]Una situazione parzialmente diversa riguarda le annotazioni 'a testo continuo'. In questo caso, è infatti possibile raccogliere informazioni sulla frequenza di occorrenza almeno per la porzione di testo completamente annotata.

connessa con le potenzialità di disporre di basi di conoscenza nelle quali l'informazione semantico–lessicale di dominio sia descritta in modo formale tale da essere utilizzata in compiti di estrazione dell'informazione ('Information Extraction') o di 'Text Mining'.

L'esempio più significativo in questo senso è quello del dominio biomedico, per il quale esiste il maggior numero di iniziative. Più che in altri domini è infatti riconosciuta la centralità di risorse che permettano di rendere esplicito il collegamento tra contenuto proposizionale e realizzazione linguistica. Come dimostrato da Cohen et al. (2008), sono centrali per lo sviluppo di sistemi di estrazione di conoscenza specialistica da testi di letteratura biomedica informazioni riguardo al modo in cui, ad esempio, a fenomeni di alternanza sintattica relativa ai comportamenti di verbi e sostantivi (es. casi di alternanza attivo/passivo o di nominalizzazione) corrisponde un comportamento semantico comune.

6.4.1 Usi nel dominio biomedico

Le iniziative volte all'uso di modelli per la rappresentazione del significato lessicale condotte in ambito biomedico sono principalmente legate al diffuso interesse per la costruzione di banche dati terminologiche e di risorse ontologiche di dominio (Bodenreider, 2006), così come per l'annotazione sintattico–semantica di corpora di letteratura biomedica.

I due principali progetti avviati in quest'ambito sono dedicati all'estensione e specializzazione di FrameNet e della PropBank. In entrambi i casi i due obiettivi principali sono *i)* quello di utilizzare un modello di organizzazione del significato per condurre uno studio di come le specificità del linguaggio biomedico influenzino l'organizzazione della semantica di dominio e *ii)* quello di utilizzare le risorse semantico–lessicali così costruite per compiti di gestione automatica della conoscenza di dominio.

A questo scopo è stato creato BioFrameNet (Dolbey, 2009), un'estensione di FrameNet che consiste nell'aggiunta di una serie di nuovi frames e di FEs specifici, evocati da LUs rappresentative per la descrizione, in particolare, del trasporto intracellulare di proteine. È il caso, ad esempio, dell'aggiunta del frame PROTEIN_TRANSPORT, che comprende i FEs 'Transport_destination', 'Transport_locations', 'Transport_origin', 'Transported_entity', evocato da *transportation.n, transport.v, export.n, migrate.v*, ecc...

Oltre che alla costruzione di una risorsa semantico–lessicale di dominio, il lavoro di Dolbey è finalizzato, in primo luogo, a dimostrare come i principi

teorici e organizzativi di FrameNet permettano di mettere in evidenza alcune particolarità grammaticali di dominio, grazie alla rappresentazione esplicita dei vincoli combinatori (sintattici e semantici) tra unità lessicali del linguaggio della biologia molecolare. A questo scopo, BioFrameNet si presenta anche come un corpus di testi annotati con informazione semantica 'a frame'.

In secondo luogo, il suo studio mostra come i frames di BioFrameNet e le classi ontologiche di un'ontologia di dominio costituiscano due modi complementari di organizzare la conoscenza. Come dimostrato anche da Uematsu et al. (2009), infatti, i principi di annotazione semantica di FrameNet consentono di creare un ponte tra il testo e la conoscenza di dominio. L'organizzazione del contenuto testuale come 'rete di Semantic Frames' permette cioè di collegare l'informazione di tipo semantico–lessicale con la conoscenza degli esperti di dominio, a sua volta organizzata in una 'rete di nodi ontologico–concettuali' (l'ontologia di dominio).

Infine, Dolbey (2009) ha indagato la possibilità di usare il BioFrameNet costruito come risorsa di riferimento in compiti di elaborazione semantica automatica del testo, quali il riconoscimento e l'annotazione automatici di ruoli semantici ('Semantic Role Labeling'). L'affidabilità di questo compito è infatti sperimentata con successo da Harabagiu e Bejan (2010), che hanno usato BioFrameNet per svolgere un compito di analisi semantica automatica ('semantic parsing') finalizzata all'estrazione di eventi che descrivono cicli biomedici.

Iniziativa analoga alla costruzione di BioFrameNet è descritta da Kokkinakis e Toporowska (2010), che hanno messo a punto una metodologia volta a estendere il FrameNet per la lingua svedese (tutt'ora in corso di sviluppo) a partire da testi rappresentativi del linguaggio medico e clinico. Anche in questo caso, l'obiettivo ultimo è quello di usare tale risorsa per costruire un sistema di estrazione automatica d'informazione.

Obiettivo del progetto PASBio (Wattarujeekrit et al., 2004) e di quello finalizzato allo sviluppo di BioProp (Chou et al., 2009) è invece quello di specializzare la PropBank con informazione relativa al comportamento sintattico–semantico di verbi contenuti in abstracts di articoli di MEDLINE. La struttura argomentale di verbi della lingua comune contenuti nella PropBank è stata infatti confrontata con quella degli stessi verbi presenti negli articoli biomedici, verificando se e come essa si modifichi.

Così, ad esempio, il verbo *to express*, che nella PropBank ha due comportamenti sintattico–semantici legati ai due sensi 'say' e 'send very quickly', in PASBIO ha il senso di 'manifest the effects of a gene or genetic trait' con la

seguente lista di argomenti: 'Arg0 : named entity (gene or gene products)', 'Arg1 : property of the existing name entity', 'Arg2 : location refering to organelle, cell or tissue'.

Tale processo ha portato alla costruzione, in modo manuale da parte di Wattarujeekrit et al. (2004) e in modo semi–automatico da parte di Chou et al. (2009), di una 'banca di proposizioni' relative al dominio della biologia molecolare. In entrambi i casi, la finalità era quella di creare le premesse per un sistema di estrazione automatica d'informazione.

Un diffuso interesse è stato inoltre dimostrato circa la possibilità di estendere il WordNet generico, specializzandolo con nuovi synsets a partire dalla selezione automatica di termini rilevanti rintracciati in corpora di testi biomedici (Buitelaar e Sacaleanu, 2002), allo scopo di migliorare i risultati dei sistemi di recupero dell'informazione utilizzati sia da esperti di dominio sia dal cittadino comune (Smith e Fellbaum, 2004).

In questo senso, è di rilievo lo studio condotto da Poprat et al. (2008), nel quale sono messi in luce gli ostacoli incontrati nel processo di estensione di WordNet al dominio biomedico. Tra gli ostacoli individuati dagli autori, molti sono dovuti a caratteristiche costitutive dell'archiettura della risorsa sviluppata per la lingua comune, caratteristiche che non permettono di fornire una rappresentazione adeguata di alcune specificità del linguaggio specialistico. È il caso ad esempio della struttura dati stessa di WordNet che impone *i)* che una parola non può avere più di 16 omonimi e non possa dunque essere parte di più di 16 synsets, limitando in questo modo la piena rappresentazione dell'ambiguità del lessico biomedico, *ii)* che una parola non può avere più di 425 caratteri, escludendo così i lunghi composti tipici del lessico biomedico, *iii)* che limitando la varietà di relazioni semantiche possibili ne esclude alcune fondamentali per il dominio.

6.4.2 Usi in altri domini

Le iniziative condotte nell'ambito di altri domini specialistici si configurano principalmente come specializzazioni ed estensioni di FrameNet e WordNet e sono strettamente collegate allo sviluppo di risorse per lingue diverse dalla lingua inglese.

Il modello FrameNet è usato nei seguenti progetti che riguardano:

- il linguaggio calcistico: come descritto da Schmidt (2008), che descrive la metodologia di annotazione semantica con informazione 'a frame' di

resoconti in inglese, tedesco e francese di partite calcistiche. Tale metodologia è finalizzata, secondo i principi di FrameNet, alla costruzione della risorsa lessicografica "Kicktionary"[57], un dizionario multilingue contenente le principali LUs evocatrici delle situazioni–tipo (frames) calcistiche più significative;

- il linguaggio ecologico–ambientale: come descritto da Reimerink et al. (2010), nel cui studio l'annotazione semantica di un corpus trilingue inglese, spagnolo, tedesco di testi in cui sono descritti eventi atmosferici, idrogeologici, ecc... è finalizzata alla costruzione di un "EcoLexicon"[58];

- il linguaggio relativo all'ambito dell'assistenza software/hardware telefonico: come descritto da Dinarelli et al. (2009), che riportano il lavoro condotto nell'ambito del progetto LUNA ("Language UNderstanding in multilingAl communication systems")[59], finalizzato allo sviluppo di un sistema avanzato di riconoscimento vocalico. In questo caso, il punto di partenza è costituito da un corpus di dialoghi in italiano, francese e polacco nei quali sono state annotate istanze di frames inerenti l'assistenza tecnica in ambito telefonico;

- il linguaggio usato nei brevetti: come descritto da Dinarelli et al. (2008), che descrivono l'utilizzo di FrameNet unicamente come modello di annotazione semantica di un corpus di brevetti nell'ambito del progetto "PATExpert"[60].

Per quanto riguarda le iniziative legate a WordNet, è qui d'interesse ricordare quella rivolta alla costruzione di "WordNet Domains"[61], risorsa creata in modo semi–automatico tramite l'aggiunta al WordNet generico di etichette che segnalano l'appartenza di un synset ad un dominio specifico (es. architettura, medicina, ecc...) (Magnini e Cavaglià, 2000). A partire dall'idea che un 'dominio' non sia altro che un insieme di parole tra le quali esistono relazioni semantiche particolarmente strette, più che nella lingua comune (Magnini et al., 2002), tale risorsa è stata utilizzata con lo scopo di migliorare i risultati di sistemi di disambiguazione di senso in corpora di dominio.

[57] http://www.kicktionary.de/index.html
[58] http://manila.ugr.es/visual/index_e.html
[59] http://www.ist-luna.eu/
[60] http://www.patexpert.org/
[61] http://wndomains.fbk.eu/index.html

Più in particolare, il WordNet generico sviluppato per la lingua inglese è stato esteso e specializzato per la lingua italiana nella costruzione di risorse lessicali specialistiche per i seguenti domini:

- architettura: per il quale è stato sviluppato, nell'ambito del progetto "ArchiWordNet" (Bentivogli et al., 2004), un thesaurus bilingue italiano/inglese di termini architettonici ed edilizi italiani per il recupero di immagini fotografiche contenute in banche dati[62];

- economia: nell'ambito del progetto "Economic–WordNet" finalizzato alla creazione di synsets di dominio;

- filosofia: per il quale, nell'ambito del progetto "Philonet"[63], sono stati sviluppati una serie di synsets relativi a concetti filosoficamente rilevanti, utilizzati per l'annotazione sematica di testi filosofici;

- navigazione e commercio marittimo: finalizzato alla costruzione di un database semantico di terminologia marittima (Marinelli et al., 2004)[64].

6.4.3 Usi nel dominio giuridico

In confronto alla vasta gamma di iniziative condotte per il dominio biomedico, nel dominio giuridico le attività rivolte alla rappresentazione esplicita e strutturata dell'informazione semantico–lessicale di dominio sono sino ad oggi ancora relativamente poche. Come precedentemente discusso, tale ritardo è riconducibile da un lato alla natura della lingua del diritto, che, strettamente intrecciata con la lingua comune, pone problemi di analisi automatica diversi da quelli rappresentati dal linguaggio biomedico; dall'altro, esso è ascrivibile alla natura stessa della materia giuridica, così legata ad un processo individuale di interpretazione del testo da essere difficilmente rappresentabile in modo condiviso da tutti gli esperti di dominio.

[62] ArchiWordNet, Economic-WordNet, Philonet e WordNet Domains sono realizzati nell'ambito delle attività legate a MultiWordNet (Pianta et al., 2002), progetto avviato dalla Fondazione Bruno Kessler e finalizzato alla costruzione di un WordNet multilingue.

[63] http://www.nyu.edu/its/humanities/ach_allc2001/posters/bentivogli/index.html

[64] Tale estensione di dominio è stata condotta nell'ambito delle attività di costruzione di ItalWordNet (Roventini et al., 2000), progetto realizzato presso l'Istituto di Linguistica Computazionale del CNR di Pisa e finalizzato alla creazione di un WordNet per la lingua italiana all'interno del progetto EuroWordNet.

Tali caratteristiche del dominio giuridico sono considerate in questo lavoro all'origine del motivo per cui i modelli di rappresentazione del significato sono per lo più usati come modelli per l'annotazione semantica di testi giuridici, più che per la costruzione di una risorsa semantico–lessicale di riferimento per il dominio. Come descritto in quanto segue, l'unica eccezione è rappresentata dalla risorsa JurWordNet. Il considerare infatti un lessico di dominio come una sistematizzazione universalmente condivisa della conoscenza semantico–lessicale del dominio è uno dei principali ostacoli. Ciò è strettamente collegato alla grande varietà di ontologie giuridiche[65]. Tale varietà, mentre nel dominio biomedico è attribuibile alla gamma di sottodomini interessati (organizzazione delle relazioni tra geni e proteine, tra le varie entità chimiche, ecc...), è invece in questo caso riconducibile alla mancanza di una visione condivisa su come organizzare sia i concetti fondamentali del diritto (quelli contenuti nelle cosiddette 'core ontologies') sia quelli relativi ai domini oggetto del diritto (quelli cioè contenuti nelle 'domain–specific ontologies').

Più ampia è al contrario la gamma di studi finalizzati all'annotazione semantica di collezioni documentali giuridiche. Mentre una descrizione dei vari approcci sino ad oggi messi a punto e basati sull'uso di strumenti di Trattamento Automatico del Linguaggio è riportata nel Paragrafo 2.3.2.2, in quanto segue sono passati in rassegna i lavori di chi utilizza i modelli di rappresentazione del significato descritti in questo capitolo per rendere esplicita l'informazione semantico–lessicale di dominio. È questo il caso di Rathert (2006) e di Mustafaraj et al. (2006), che hanno adottato FrameNet come modello di riferimento, e di Wyner e Peters (2010b), che hanno utilizzato il lessico VerbNet all'interno della loro metodologia di annotazione semantica di un corpus di sentenze in lingua inglese.

Nel primo caso i principi teorici e organizzativi di FrameNet sono utilizzati da Rathert (2006) per sperimentare come la comprensibilità di una sentenza in lingua tedesca possa essere empiricamente stabilita grazie al processo di annotazione semantica del testo, attraverso la verifica *i)* della corretta realizzazione nel testo dei FEs necessari alla descrizione dei vari frames e *ii)* della corretta ricostruzione della rete di frames (delle relazioni 'frame–to–frame') all'interno dell'intero testo.

Mustafaraj et al. (2006), invece, hanno inserito la fase di annotazione

[65]Vedi Casellas (2011) per un ricco e aggiornato stato dell'arte sulle diverse ontologie giuridiche sviluppate sino ad oggi.

semantica 'a frame' condotta sul modello FrameNet in una completa catena di analisi del testo che, a partire da una fase di annotazione sintattica automatica, culmina con il rendere espliciti i ruoli semantici ricoperti dai vari partecipanti ai principali frames presenti in un corpus composto da decisioni contenute in sentenze in lingua tedesca. In questo caso, la finalità è quella di costruire una base di conoscenza per sviluppare un sistema *i)* di recupero automatico di sentenze e *ii)* di 'legal reasoning', a partire dall'identificazione del contenuto di decisioni giudiziarie.

L'acquisizione automatica di informazione relativa a 'fatti' e 'soggetti coinvolti' presenti in un corpus di sentenze in lingua inglese è l'obiettivo perseguito da Wyner e Peters (2010b). In questo caso, gli autori arricchiscono l'approccio messo a punto da Wyner (2010) e Wyner e Peters (2010a) con l'informazione semantico–lessicale già codificata in VerbNet. Basando infatti la loro metodologia di annotazione semantica sull'identificazione nel testo della struttura predicato–argomenti di alcuni dei verbi più significativi, riescono a superare l'ostacolo posto, in fase di estrazione dell'informazione, dalla varietà di realizzazione sintattica nel testo di 'fatti' (linguisticamente realizzati come predicati) e 'soggetti coinvolti' (gli argomenti). La finalità ultima è quella di fornire un utile ausilio al giudice impegnato nel recupero dei precedenti giudiziari rilevanti per formulare la decisione finale.

6.4.3.1 JurWordNet

Il progetto[66], realizzato congiuntamente dall'Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG–CNR) di Firenze e dall'Istituto di Linguistica Computazionale (ILC–CNR) di Pisa, nasce, come chiarito da Sagri (2002), con un duplice esplicito intento: quello di creare cioè una risorsa semantico–lessicale (un lessico semantico) per il diritto, fondata sui principi organizzativi di WordNet, e quello di offrire un modello di descrizione strutturata della conoscenza di dominio.

Pertanto, JurWordNet da un lato estende e specializza ItalWordNet, attraverso la creazione di synsets di dominio collegati con il WordNet italiano tramite un procedimento di 'plug–in' (Bertagna et al., 2004), il quale stabilisce una serie di relazioni tra i synsets di ItalWordNet e quelli di JurWordNet. Si tratta per lo più di relazioni che permettono di rendere espliciti i modi del

[66]http://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/JurWordNetWordNetPerIlDiritto.htm

riuso specialistico del lessico comune, come nel caso di *i)* termini che, in ambito giuridico, hanno bisogno di specificazioni tecniche di significato, come ad esempiio *autorizzazione*, unico termine del synset {autorizzazione}, definito dalla glossa come "atto con cui un privato o l'autorità pubblica permette a un soggetto l'esercizio di un diritto o di una facoltà" e non genericamente 'l'atto di rendere qualcosa legittimo', o di *ii)* termini che assumono un significato tecnico in ambito giuridico, come ad esempio *mora*, che appartiene al synset {mora}, definito dalla glossa come "ritardo ingiustificato nell'adempimento di un'obbligazione dal quale può derivare l'obbligo del risarcimento del danno".

Dall'altro, JurWordNet si presenta come un'ontologia linguistica grazie al collegamento tra i synsets di livello gerarchico superiore e le classi della "Core Legal Ontology" (CLO)[67] (Gangemi et al., 2005) nella quale sono organizzati i concetti fondamentali ('core') della realtà giuridica, condivisi da ordinamenti e sistemi giuridici diversi (es. 'soggetto giuridico', 'evento e atto giuridico', 'violazione', ecc...) (Gangemi et al., 2003b). Ad esempio, il synset {funzione, funzione giuridica}[68] è legato alla classe CLO #LegalFunction[69].

Questa doppia veste permette di rendere esplicite non solo le relazioni **semantiche** tra concetti (synsets) giuridici, stabilite sulla base dei rapporti paradigmatici di significato tra i termini del lessico giuridico, ma anche quelle **ontologiche**, definite in base alla natura ontologica delle entità fondamentali del mondo giuridico. Così, ad esempio, da un lato, la rete semantica tra synsets mette in collegamento il synset {funzione, funzione giuridica} con i synsets {Pubblico Ministero} e {ufficio} sulla base di una relazione di iponimia; dall'altro, sulla base del collegamento tra i synsets e le classi della CLO, viene reso esplicito che le unità lessicali parte del synset {funzione, funzione giuridica} sono istanze lessicali del concetto #LegalFunction, a sua volta legato dalla relazione ontologica 'subClassOf' alla classe #AgentRole[70].

Tale approccio è particolarmente efficace in un contesto di multilinguismo giuridico, dove l'armonizzazione della terminologia giuridica tra lingue diverse avviene a livello di condivisione di concetti piuttosto che a livello ter-

[67] http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl

[68] Il synset è definito dalla glossa: "potestà esecutiva per un interesse non proprio ma oggettivo o di altri".

[69] Nella CLO la classe è così definita: "Legal functions are legal roles, only played by legal subjects. Among legal functions, so–called Primary Functions (e.g. Son, Heir, Citizen) are defined by constitutive norms".

[70] Nella CLO la classe è così definita: "A Role that classifies an Agent".

minologico. Per questo motivo, a partire da JurWordNet è stato sviluppato nell'ambito del progetto LOIS ("Legal Ontologies for Knowledge Sharing") (Tiscornia, 2007) un lessico giuridico multilingue, unendo, tramite relazioni di equivalenza, i WordNets di EuroWordNet sviluppati per la lingua di sei diversi paesi (Germania, Olanda, Inghilterra, Portogallo, Repubblica Ceca e Italia).

Coerentemente con questa doppia natura della sua architettura, JurWordNet è stato pensato per essere usato in compiti applicativi diversi che vanno dall'annotazione semantica di testi legislativi, grazie all'utilizzo dei synsets come metadati semantici informativi del profilo semantico–funzionale (del 'disposto') di una legge[71], all'uso in compiti di gestione dell'informazione a partire dalla base di conoscenza giuridica così costruita.

È qui inoltre d'interesse mettere l'accento su due aspetti centrali di come l'informazione semantico–lessicale giuridica è organizzata in JurWordNet. Sebbene la risorsa sia strutturata sulla base del modello WordNet, tuttavia, in quanto estensione di dominio di ItalWordNet, in essa la rete di synsets è definita non solo dalle relazioni previste nel WordNet di Princeton, ma anche da quelle contenute in EuroWordNet[72]. Pertanto, i synsets di JurWordNet sono legati anche da relazioni di tipo 'role_agent', 'role_result', 'involved', ecc...[73].

Ad esempio, il synset {assunzione}[74] è legato da una relazione *i)* di iperonimina al synset {atto giuridico}[75]; *ii)* di iponiminia ai synsets {assunzione prove}, {assunzione obbligatoria}, {assunzione temporanea}, {assunzione in prova}, {assunzione straordinaria}, {assunzione diretta}, ecc...; e *iii)* di 'role_result' al synset {assumere}[76]. Quest'ultima è un tipo di relazione di Euro-

[71]Tale applicazione è stata esplicitamente pensata nel contesto del progetto nazionale "Norme in rete" (vedi Paragrafo 2.3.1), con lo scopo di fornire una fonte di metadati 'semantici' oltre a quelli relativi al profilo formale dell'articolato di un testo legislativo.

[72]Come ricordato nel Paragrafo 6.4.2, ItalWordNet rappresenta la sezione italiana di EuroWordNet. Di conseguenza, le relazioni tra synsets sono quelle previste in EuroWordNet, parzialmente diverse da quelle del WordNet di Princeton e non sempre sovrappronibili, come dimostrato da Pazienza et al. (2008).

[73]Per una lista completa delle relazioni contenute in JurWordNet vedi http://godel.ittig.cnr.it/jwn/web/relazioni.php

[74]Il synset è definito dalla glossa: "atto che dà inizio a un rapporto di lavoro subordinato".

[75]Il synset è definito dalla glossa: "qualsiasi azione compiuta volontariamente che implica conseguenze giuridiche".

[76]Il synset è definito dalla glossa: "prendere presso di sé con contratto di assunzione,

WordNet che si può instaurare anche tra synsets di categorie morfosintattiche diverse (quali sostantivo e verbo)[77].

Inoltre, la rete di relazioni in JurWordNet rende espliciti anche i rapporti semantico–lessicali specifici di dominio, quelli cioè propri della teoria generale del diritto che si intende organizzare in modo strutturato. Così, ad esempio, il synset {persona giuridica}[78] è legato da una relazione di iponimia non solo ai synsets {persone giuridiche private} e {persone giuridiche pubbliche}, ma anche al synset {banca}[79] in quanto 'soggetto giuridico'.

Proprio in seguito all'adozione di questi criteri specifici di dominio, molti dei synsets contengono una sola parola. In base alla definizione di sinonimia del modello WordNet, sono rari infatti i casi di parole del lessico giuridico che possono essere liberamente sostituite in una frase senza alterarne l'accettabilità[80].

6.5 Le potenzialità di FrameNet per l'annotazione semantica di testi giuridici

I confronti tra FrameNet e Wordnet[81], da un lato, e tra FrameNet e gli altri progetti basati sulla rappresentazione a livello paradigmatico del significato[82], dall'altro, hanno sin qui permesso di mettere in luce come varie caratteristiche specifiche di FrameNet contribuiscano a renderlo un modello di rappresentazione e organizzazione del significato particolarmente espressivo per la descrizione del contenuto di testi giuridici.

A conclusione di questo capitolo, l'obiettivo è ora di focalizzare l'attenzione sulle potenzialità di adottare FrameNet come modello di riferimento per l'annotazione semantica. I vantaggi sono principalmente di due tipi.

In primo luogo, come discusso nel Paragrafo 6.5.1, i principi organizzativi di FrameNet, basati su un'organizzazione a livello **sintagmatico** del

prendere alle proprie dipendenze".

[77]La relazione inversa, 'involved_result', tra i synsets {assumere} e {assunzione}, consente di legare il synset verbale a quello deverbale.

[78]Il synset è definito dalla glossa: "soggetto di diritto diverso dalla persona fisica".

[79]Il synset è definito dalla glossa: "È quell'impresa autorizzata all'esercizio dell'attività bancaria ossia alla raccolta del risparmio tra il pubblico e all'esercizio del credito".

[80]Vedi Paragrafo 6.2.1.

[81]Vedi Paragrafo 6.2.2.

[82]Vedi Paragrafo 6.3.4.

significato, consentono di rendere espliciti aspetti di descrizione del significato di un testo giuridico complementari a quelli offerti da JurWordNet. In questo modo, dunque, una metodologia di annotazione semantica del testo basata su tali principi permette di catturare elementi del contenuto del discorso giuridico nuovi rispetto a quelli noti grazie alla risorsa di dominio oggi esistente.

In secondo luogo, la centralità attribuita dal modello FrameNet all'**uso linguistico** come punto di partenza da cui ha origine l'intero processo di comprensione semantica del testo è un elemento chiave. Esso consente infatti di inserire la metodologia di annotazione semantica proposta in questo studio all'interno del dibattito teorico degli studi in filosofia analitica del diritto circa l'importanza dell''uso' dei concetti giuridici e delle 'regole d'uso linguistico' proprie del discorso giuridico[83]. In questo senso, come discusso nel Paragrafo 6.5.2, la nozione di 'Semantic Frame' intesa come schematizzazione di un contesto conoscitivo–tipo può consentire di risolvere alcuni problemi di rappresentazione della conoscenza giuridica ben noti e riconosciuti nella comunità di ricerca in AI&Law ma lasciati sino ad oggi per lo più irrisolti.

Le discussioni condotte nei successivi paragrafi sono inoltre di particolare interesse dal momento che permettono di contestualizzare e motivare alcune delle scelte compiute in fase di definizione della metodologia di annotazione semantica. Le discussioni circa le potenzialità di FrameNet come modello di descrizione del significato e di rappresentazione della conoscenza costituiscono infatti il punto di partenza per la descrizione della strategia di annotazione esposta nel Paragrafo 7.2.

6.5.1 Aspetti di descrizione del significato

La ragione principale che ha spinto ad adottare in questo lavoro FrameNet come modello di annotazione semantica di testi giuridici è legata in primo luogo ai principi organizzativi seguiti nella descrizione del significato, principi ortogonali ma nello stesso tempo complementari a quelli di WordNet, su cui JurWordNet è modellato. Come discusso nel Paragrafo 6.2, il diverso punto di vista sul significato e i modi diversi di descriverlo sono al centro della questione.

L'aspetto di maggiore rilievo sta proprio nella scelta dell'unità minima di descrizione del significato. Dal diverso approccio teorico derivano infatti

[83] Vedi Paragrafo 2.2.

scelte diverse su come rappresentare in modo strutturato il significato: *i)* un punto univocamente determinato dalla rete di relazioni semantiche nel quale esso è inserito, in WordNet, *ii)* una struttura composizionale definita dagli elementi semantici che la compongono, in FrameNet.

Tali differenze hanno ripercussioni particolarmente evidenti nel caso della rappresentazione del significato in un dominio specialistico, dove i due approcci corrispondono a due diversi modi di rappresentazione della conoscenza di dominio: *i)* il tipo di informazione fornita da WordNet consente di 'definire' i principali concetti di dominio, contribuendo in questo modo a delimitare la rete di significati caratterizzanti un determinato dominio di conoscenza; *ii)* l'informazione relativa al comportamento combinatorio del lessico di dominio offerta da FrameNet consente di spostare il cuore della rappresentazione sulla 'descrizione' del modo in cui i principali elementi conoscitivi di dominio interagiscono tra loro.

Date queste premesse, focalizzando per ora l'attenzione su aspetti di descrizione del significato, è qui intenzione proporre un esempio dimostrativo di come una metodologia di annotazione semantica di testi giuridici basata sul modello FrameNet possa essere considerata una strategia complementare ad una basata sul modello WordNet, come quella offerta da JurWordNet.

L'esempio è dato dall'annotazione semantica del seguente periodo, estratto dal corpus AMBnorm(Stato):

a) *In caso di mancato rispetto del programma di cui al comma 4, ovvero di mancata segnalazione ai sensi del comma 2, il soggetto gestore ha l'**obbligo** di risarcire i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori.*

L'attenzione è qui posta sulla rappresentazione del significato del sostantivo *obbligo*. In JurWordNet la parola è parte dei due synsets {obbligo}[1] e {obbligo}[2][84]. Pertanto, il significato di *obbligo* deve essere 'disambiguato', deve cioè essere stabilito a quale dei due synsets esso appartenga.

Il processo di annotazione semantica consiste dunque nel rendere esplicito il senso del sostantivo *obbligo*, sulla base *i)* delle definizioni dei due synsets candidati e *ii)* delle relazioni semantiche che li legano ad altri synsets in JurWordNet. Come si può vedere nella Tabella 6.1, dove per ciascuno dei due synsets sono riportate le definizioni (prima colonna), il tipo di relazione semantica (seconda colonna) e i synsets collegati (terza colonna), la definizione

[84]In entrambi i casi i synsets sono costituiti da una sola parola.

che più si adatta alla descrizione del significato della parola *obbligo* nel periodo a) è quella del synset {obbligo}[1], legato da una relazione di iponimia a 47 synsets che costituiscono delle fattispecie di comportamenti imposti.

Definizione	Relazione	Synset collegato
{obbligo}[1]		
"vincolo a tenere un certo comportamento, a fare o non fare una determinata cosa, derivante dal rispetto di una norma morale, religiosa o giuridica"	has_hyponym	47 synsets[85]
{obbligo}[2]		
"dovere cui è tenuto ad adempiere il soggetto passivo di un'obbligazione"	has_hyperonym	{dovere giuridico}[1]
	near_synonym	{onere}[1]

Tabella 6.1: I synsets in JurWordNet che descrivono il significato del sostantivo *obbligo*.

L'obbligo di cui si parla nel periodo a) è infatti uno dei possibili comportamenti che un soggetto può essere obbligato a tenere. A conferma di ciò, è da notare che il synset {obbligo risarcimento} è tra i synsets iponimi di {obbligo}[1].

L'informazione che non è però possibile ricavare dal periodo a) sulla base dell'annotazione semantica basata su JurWordNet è quella relativa ai partecipanti l''obbligo di risarcimento'. Supplisce a questa mancanza il tipo di rappresentazione del significato di FrameNet. Sulla base infatti dello schema

[85] I synsets sono: {obbligo informazione}, {obbligo soggiorno}, {collaborazione}[3], {obbligo a contrarre}, {obbligo pubblicazione}, **{obbligo risarcimento}**, {obbligo dipendente}, {obbligo tributo}, {obbligo contributivo}, {obbligo garanzia}, {obbligo rendiconto}, {obbligo alimentare}, {obbligo vigilanza}, {obbligo convenzionale}, {obbligo segnalazione}, {obbligo denunciare}, {obbligo rilascio}, {obbligo di dimora}, {obbligo adempimento}, {obbligo giuridico}, {obbligo pagamento}, {obbligo contributo}, {obbligo orario}, {obbligo conducente}, {obbligo presenza}, {obbligo fedeltà}, {obbligo dimorare}, {obbligo stato}, {obbligo contrattuale}, {obbligo esecuzione}, {obbligo registrazione}, {obbligo datore}, {obbligo versamento}, {obbligo pagare}, {obbligo assicurazione}, {obbligo giudice}, {obbligo retribuzione}, {obbligo assicurativo}, {obbligo alimento}, {obbligo mantenimento}, {obbligo notificazione}, {obbligo locatore}, {obbligo rispetto}, {obbligo comunicazione}, {obbligo politico}, {obbligo dichiarativo}, {obbligo patente}.

di annotazione adottato in FrameNet, è possibile rendere esplicita l'informazione relativa al 'tipo di situazione' evocata dal sostantivo *obbligo*, al 'soggetto vincolato' a tenere un determinato tipo di comportamento, alle 'circostanze' che portano il soggetto a dover tenere un determinato comportamento.

Il periodo a) è dunque semanticamente annotato come segue:

a.1) [BEING_OBLIGATED] [*In caso di mancato rispetto del programma di cui al comma 4, ovvero di mancata segnalazione ai sensi del comma 2,* $_{Condition}$] [*il soggetto gestore* $_{Responsible\_party}$] [*ha* $_{Supp}$] *l'**obbligo*** [*di risarcire i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori* $_{Duty}$].

Un'annotazione tale permette così di rendere esplicito che il sostantivo *obbligo* (LU segnalata in grassetto) è in grado di evocare in questo periodo il frame BEING_OBLIGATED, che descrive la situazione–tipo in cui "Under some Condition, usually left implicit, a Responsible_party is required to perform some Duty. If they do not perform the Duty, there may be some undesirable Consequence, which may or may not be stated overtly".

Oltre alla definizione del senso di *obbligo* offerta da JurWordNet, è in questo modo possibile ricostruire gli elementi conoscitivi (i FEs il cui nome è riportato a pedice) che lo costituiscono, insieme alle relative istanze particolari (la realizzazione lessicale dei FEs racchiusa tra parentesi quadre).

Inoltre, la fattispecie dell'obbligo, resa esplicita dal synset {obbligo di risarcimento} di JurWordNet, è rappresentata rintracciando in FrameNet un frame in grado di descrivere il tipo di dovere imposto. Si tratta in questo caso della situazione descritta dal frame FINING[86]. Arricchito con questa informazione, il periodo a.1) è annotato come segue:

a.2) [FINING] *In caso di mancato rispetto del programma di cui al comma 4, ovvero di mancata segnalazione ai sensi del comma 2, il soggetto gestore ha l'obbligo di **risarcire*** [*i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori* $_{Reason}$]. [CNI $_{Payer}$][87]

[86]Il frame è così definito in FrameNet: "The Payer is (legally) forced to pay a Fine by an official Speaker as a punishment for some action (the Reason). The Speaker represents an entity which receives the payment."

[87]Si tratta di un caso di 'Constructional Null Instantiation' per cui il FE non è lessicalmente istanziato a causa della struttura a soggetto controllato della frase argomentale *di risarcire i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori.*

In conformità con i principi organizzativi di FrameNet, questo tipo di rappresentazione del significato è condotto sulla base della realizzazione sintattica degli elementi conoscitivi oggetto di attenzione. Come recentemente ricordato da Fillmore e Baker (2010), infatti, obiettivo di FrameNet è quello di "assemble information about alternative ways of expressing concepts in the same conceptual domain", attraverso la descrizione delle proprietà combinatorie (a livello sintattico e semantico) di tutte le parole che esprimono un determinato concetto. In generale, tale approccio alla descrizione del significato "makes it possible to separate the notion of the conceptual underpinnings of a concept from the precise way in which the words anchored in them get used" (Fillmore e Atkins, 1992).

Tali criteri di descrizione del significato fanno di FrameNet uno strumento particolarmente versatile per la descrizione delle principali specificità sintattiche di un testo. Ciò permette infatti di porre particolare attenzione al modo in cui alcune delle caratteristiche specifiche della lingua del diritto si fanno veicolo della semantica del discorso giuridico. Nel periodo a) annotato e discusso in questo paragrafo, ad esempio, il significato del sostantivo *obbligo* è rappresentato a partire dalla costruzione a verbo supporto nella quale il sostantivo occorre. Di conseguenza, il 'soggetto obbligato ad adempiere l'obbligo' (FE Responsible_party) è sintatticamente realizzato come il soggetto del verbo supporto *avere*, il 'dovere' (FE Duty) è realizzato come la frase argomentale dipendente dal sostantivo *obbligo*, ecc...

È qui d'interesse infine ricordare che un ulteriore vantaggio di FrameNet riguarda l'accento posto sul carattere composizionale della descrizione del significato. Alla luce infatti dei principi della 'Frame Semantics Theory', il significato di una parola è pienamente descritto solo attraverso un graduale processo interpretativo, un processo di progressiva ricostruzione degli elementi sintattici e semantici fondamentali alla completa comprensione della situazione conoscitiva evocata.

Questi principi sono visti in questo lavoro in linea con quanto messo in evidenza negli studi finalizzati a definire il livello di comprensibilià testuale. Come ricordato, in particolare, da Piemontese (2001, p. 128), "si può parlare di chiarezza, semplicità e precisione dei testi solo se un testo, oltre che leggibile, è anche comprensibile, cioè costruito dal punto di vista logico–concettuale in modo controllato". In questo senso FrameNet è qui visto come uno strumento indispensabile per rintracciare quella che Garavelli (2001, p. 176) ha definito come una delle "qualità essenziali, irrinunciabili" di un testo, cioè "la 'buona formazione' della struttura argomentativa".

6.5.2 Aspetti di rappresentazione della conoscenza

Il principio di organizzare la conoscenza di dominio a livello sintagmatico fornendo una rappresentazione delle entità rilevanti sulla base della loro struttura interna piuttosto che sulla base della loro organizzazione tassonomica è un approccio alla rappresentazione della conoscenza considerato particolarmente espressivo nel campo dell'intelligenza artificiale (Minsky, 1975). È il caso dei linguaggi di rappresentazione della conoscenza cosiddetti 'frame–based' o 'object–oriented', basati sulla descrizione prototipica delle entità (oggetti) da rappresentare i quali vengono scomposti in singoli elementi costitutivi ('slots'), organizzati appunto in una struttura che li sussume ('frame')[88].

Da un punto di vista applicativo, le maggiori potenzialità nello sviluppo di sistemi di organizzazione della conoscenza dotati di un'architettura 'frame–based' sono evidenti soprattutto nel caso di sistemi costruiti per rappresentare in maniera strutturata conoscenza di dominio. Più che in altri casi infatti i requisiti fondamentali di 'comprensibilità, accessibilità, espressività' del sistema sono caratteristiche necessarie per consentire una chiara rappresentazione di un dominio di conoscenza regolato da strutture complesse che per essere pienamente comprese devono essere scomposte in elementi conoscitivi minimi. Come discusso da Noy et al. (2002), ad esempio, per modellare la conoscenza relativa all'anatomia umana, non è sufficiente che, per esempio, le diverse parti di un muscolo siano organizzate in gerarchie topologiche; è necessario specificare le relazioni che le legano e che ne consentono una descrizione in quanto elementi costitutivi della struttura conoscitiva 'muscolo'.

A partire da tali considerazioni è intenzione in questo lavoro mettere in luce le potenzialità di FrameNet come modello di rappresentazione e organizzazione delle principali strutture conoscitive contenute nel discorso giuridico. L'approccio 'frame–based' adottato va infatti incontro alle necessità sollevate nell'ambito della comunità in AI&Law in materia di organizzazione strutturata e computabile della conoscenza giuridica.

Recentemente Breuker (2009), riflettendo sul contributo che le ontologie giuridiche e i lessici semantici (sino ad oggi sviluppati) possono dare alla realizzazione di compiti di gestione dell'informazione giuridica, è arrivato infatti alla conclusione che nè le une nè gli altri costituiscono strumenti in grado di modellare in modo soddisfacente 'fatti e eventi giuridici'. I prin-

[88] "A frame is a data–structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party" (Minsky, 1975).

cipi organizzativi sui quali tali risorse sono costruite permettono infatti di avere informazioni solo sulla **semantica** dei termini, "what we know about terms", e non propriamente sui loro modi di organizzazione contestuale, ciò che costituisce nell'argomentazione di Breuker la **conoscenza**, "what terms mean in a particular context (domain, document, phrase, ...)". Ciò è dovuto al fatto che, sino ad oggi ci si è concentrati, secondo Breuker, solo sull'aspetto paradigmatico di rappresentazione del significato, trascurando quello sintagmatico. È per questo che egli auspica l'avvento di una futura generazione di modelli di organizzazione della conoscenza basati su principi di rappresentazione che permettano di catturare il significato dei termini nel contesto.

La prospettiva è in linea con la direttiva di ricerca degli studi sul lessico del diritto e di quelli in semiotica giuridica condotti nell'alveo della scuola di filosofia analitica del diritto[89]. In entrambi gli ambiti di studio è infatti riconosciuta la centralità del 'contesto d'uso' delle parole all'interno del discorso giuridico *i)* sia per stabilire lo statuto del lessico giuridico sulla base dell'analisi della realtà giuridica ed extragiuridica a cui i termini fanno riferimento[90] *ii)* sia per riuscire ad interpretare con successo il contenuto di un testo giuridico a partire dalla sua analisi sintattica e semantica[91].

È inoltre qui d'interesse ricordare che un tale approccio 'frame–based' era già stato adottato nella costruzione della "Frame–Based Ontology of Law" (FBO) sviluppata da van Kralingen (1997) negli anni '90. Sfortunatamente il modello proposto da van Kralingen non ha avuto molto successo nel panorama delle ontologie giuridiche ed è per questo stato abbandonato. Nell'ambito di questo lavoro è ritenuto al contrario particolarmente rilevante. Esso costituisce infatti un'importante giustificazione dell'adozione di FrameNet come modello di organizzazione della conoscenza giuridica. In quanto segue se ne riportano pertanto gli elementi fondamentali.

La FBO è basata sul principio fondamentale per cui una 'norma', elemento principe di ogni sistema giuridico, non è altro che uno 'schema di interpretazione' di uno standard di comportamento[92], formalizzabile sotto forma di

[89] Vedi Paragrafo 2.2.

[90] Vedi Paragrafo 2.2.1 e in particolare le discussioni di Belvedere (1994a).

[91] Vedi in particolare le dicussioni di Jori e Pintore (1995) relative alla riconosciuta "priorità alla risoluzione dei problemi sintattici su quelli semantici".

[92] "A norm must convey information to fulfil its function of comunicating standards of behavior. The way in which one is expected to behave must be clear from the norm. Thus, a norm serves as a scheme of interpretation" (van Kralingen, 1997).

una struttura dati (un 'frame') che sussume una serie di elementi costitutivi la norma (riportati nella Tabella 6.2), come ad esempio il destinatario della norma (Subject), le condizioni di applicazione della norma (Conditions of application), ecc... La norma si configura pertanto come un insieme di elementi che descrivono una struttura stereotipata ('norm frame'), elementi legati da una serie di relazioni 'di dominio' in base alle quali "the norm subject is commanded, prohibited, permitted or empowered (legal modality) to perform an act (act description)".

Elemento	**Descrizione**
Norm identifier	The norm identifier (used as a point of reference for the norm).
Norm type	The norm type (norm of conduct or norm of competence).
Promulgation	The promulgation (the source of the norm).
Scope	The scope (the range of application of the norm).
Conditions of application	The conditions of application (the circumstances under which a norm is applicable).
Subject	The norm subject (the person or persons to whom the norm is addressed).
Legal modality	The legal modality (ought, ought not, may, or can).
Act identifier	The act identifier (used as a reference to a separate act description).

Tabella 6.2: Gli elementi costitutivi di una norma ('norm frame') secondo la FBO descritta da van Kralingen (1997).

Parte integrante del 'norm frame' è l'atto regolato, oggetto della norma (Act identifier), a sua volta scomponibile in una serie di elementi (riportati nella Tabella 6.3) e per questo rappresentabile anch'esso come un frame ('act frame'). Tali elementi contribuiscono a descrivere diversi aspetti dell'atto regolato, come ad esempio, chi svolge l'azione regolata dalla norma (Agent), quando (Temporal aspects), in che luogo (Spatial aspects), ecc...

Di fatto, dunque, la FBO offre un modello per la rappresentazione della realtà sia giuridica (tramite il 'norm frame') sia extragiuridica (tramite l' 'act frame') contenuta in un testo. A dimostrazione di ciò, tale modello è stato applicato da van Kralingen et al. (1993) per modellare la conoscenza contenuta in alcuni articoli del codice penale olandese. In quell'occasione gli autori hanno dimostrato *i)* che le istanze dei vari elementi costitutivi di un 'norm frame' regolativo erano a loro volta formalizzabili come 'act frames' e *ii)* come gli elementi costitutivi di un 'act frame', rappresentativo ad esempio dell'atto del rubare, fossero a loro volta linguisticamente realizzati nel testo.

Elemento	Descrizione
Act identifier	The act identifier (used as a point of reference for the act).
Promulgation	The promulgation (the source of the act description).
Scope	The scope (the range of application of the act description).
Agent	The agent (an individual, a set of individuals, an aggregate or a conglomerate).
Act type	The act type. Both basic acts and acts specified elsewhere can be used.
Means	The modality of means (material objects used in the act or more specific descriptions of the act).
Manner	The modality of manner (the way in which the act has been performed).
Temporal aspects	The temporal aspects (an absolute time specification).
Spatial aspects	The spatial aspects (a specification of the location where the act takes place).
Circumstances	The circumstantial aspects (a description of the circumstances under which the act takes place).
Cause	The cause for the action (a specification of the reason(s) to perform an action).
Aim	The aim of an action (the goal visualized by the agent).
Intentionality	The intentionality of an action (the state of mind of the agent).
Final state	The final state (the results and consequences of an action).

Tabella 6.3: Gli elementi costitutivi di ogni atto regolato da una norma ('act frame') secondo la FBO descritta da van Kralingen (1997).

Infine, come dimostrato nel Capitolo 7, l'adozione dei principi organizzativi di FrameNet come modello di riferimento per l'annotazione della conoscenza contenuta in testi giuridici risulta essere particolarmente espressivo per una serie di caratteristiche specifiche.

Primo tra tutti il fatto che il 'Semantic Frame', essendo sia uno strumento di descrizione linguistica sia la chiave di accesso alla comprensione del modo in cui si struttura il contenuto proposizionale di un enunciato, costituisce un approccio alla rappresentazione della conoscenza che mette al centro il **testo**. In questo senso dunque l'annotazione semantica si configura come un processo ponte che mette in collegamento l'informazione linguistica di un testo e la conoscenza in esso contenuta.

Un ulteriore vantaggio di FrameNet riguarda il modello 'frame–based' di organizzazione della conoscenza adottato, in base al quale più frames tra loro collegati costituiscono dei 'frame–systems', nei quali "the different frames [...]

describe the scene from different viewpoints, and the transformations between one frame and another represent the effects of moving from place to place" (Minsky, 1975). Ciò consente una descrizione della conoscenza di dominio più articolata di quella generalmente proposta dalle ontologie giuridiche[93].

[93]Vedi a questo proposito le discussioni del Paragrafo 7.5.

Capitolo 7

Un caso di studio: l'annotazione semantica di scenari deontici in atti normativi statali

L'obiettivo di questo capitolo è quello di offrire una dimostrazione di come il modello FrameNet possa essere concretamente applicato come schema di annotazione semantica di testi giuridici. A questo scopo è stato scelto come caso di studio quello della rappresentazione dell'informazione relativa a **scenari deontici** presenti in atti normativi.

La motivazione di questa scelta è triplice. In primo luogo, essa è legata alla centralità dei termini espressione delle modalità deontiche in quanto veicoli della realtà giuridica contenuta in un testo. Come riconosciuto nell'ambito degli studi sul lessico giuridico condotti dai filosofi del diritto[1], sono infatti questi quei termini "attraverso i quali si esprime sul piano linguistico la funzione prescrittiva delle norme, che qualificano giuridicamente comportamenti o attribuiscono posizioni giuridiche" (Belvedere, 1994a, p. 23). Essi rappresentano cioè la realizzazione lessicale dei "Concetti Giuridici Fondamentali" definiti dalla teoria generale del diritto e strutturati in modo formale nelle cosidette 'core legal ontologies' sulla base di formalismi sviluppati nelle ricerche in AI&Law (Sartor, 2006).

[1] Vedi Paragrafo 2.2.1.

L'annotazione delle modalità deontiche costituisce per questo motivo un buon banco di prova per confrontare un approccio alla rappresentazione della conoscenza giuridica basato su principi di annotazione semantico–lessicale del testo e un approccio esclusivamente basato su presupposti teorici di organizzazione dei concetti deontici, verificandone somiglianze e differenze.

Infine, gli scenari deontici rappresentano un punto di osservazione ottimale per lo studio di come realtà **giuridica** ed **extragiuridica** si intreccino nel discorso giuridico[2]. Essi costituiscono, cioè, una buona prospettiva da cui osservare come situazioni relative alla prescrizione di comportamenti (situazioni appartenenti alla realtà 'giuridica') si leghino a situazioni del mondo (situazioni 'extragiuridiche').

L'interesse verso gli scenari deontici nasce infine dalla consapevolezza che in ambito linguistico–computazionale pochi lavori sono stati dedicati allo studio dei concetti deontici a partire dall'analisi delle loro strutture linguistiche. Un'eccezione significativa è rappresentata dallo studio interdisciplinare di Wyner (2008) che affronta questioni di logica deontica a partire dall'analisi sintattica e semantico–lessicale degli operatori deontici (es. *ought, obliged*) espressione dei principali concetti deontici (es. 'obligation'). Parte del lavoro di Wyner, finalizzato a fornire una metodologia di rappresentazione formale del 'contratto', inteso come una serie di azioni di natura deontica riconducibili a casi di 'violazione' e 'adempimento' di un obbligo, è dedicato all'analisi del rapporto tra struttura linguistica e forma logica di una proposizione.

L'obiettivo di questo capitolo è tuttavia diverso. Focalizzato sull'annotazione semantica del testo, esso mira infatti a dimostrare come la metodologia messa a punto in questo lavoro consenta di rendere esplicita l'informazione semantico–lessicale relativa agli scenari deontici attraverso un processo di annotazione linguistica stratificata del testo, il cui punto di partenza è costituito dalla fase di annotazione sintattica.

A questo scopo, a seguito della descrizione dei frames esistenti in FrameNet riconducibili alle tre modalità di 'obbligo', 'divieto' e 'permesso' (Paragrafo 7.1) e dei criteri di annotazione semantica della struttura sintattica a dipendenze del testo (Paragrafo 7.2), sono riportate e discusse alcune delle annotazioni condotte nel corpus AMBnorm(Stato). La scelta del corpus ha permesso di mantenere uniforme l'analisi semantica rispetto alla tipologia di testo giuridico e all'autorità emittente.

[2]Sulla base della distinzione operata da Belvedere (1994a).

Le discussioni che seguono sono organizzate nei paragrafi successivi in modo tale da mettere in evidenza come i principi organizzativi di FrameNet siano particolarmente adatti a rendere esplicito lo stretto legame tra informazione linguistica e di dominio. L'attenzione è posta in particolare su:

- come le due modalità di annotazione previste dal progetto FrameNet consentano di rendere esplicita l'organizzazione del contenuto proposizionale di un periodo sia per interessi di descrizione lessicografica (grazie alla modalità lessicografica) sia per scopi di completa rappresentazione degli elementi conoscitivi necessari per la piena comprensione del contenuto testuale (grazie alla modalità 'a testo continuo'). Ponendosi come una via di mezzo tra questi due approcci, inoltre, la nuova modalità di annotazione messa a punto in questo lavoro permette di accedere ad un tipo di informazione particolarmente utile in questo dominio (Paragrafo 7.3);

- come la visione del processo di accesso al significato testuale, come un processo di progressiva esplicitazione di diversi livelli di conoscenza (sintattica e semantica) stratificati nel testo, permetta di mettere in luce il modo in cui la semantica del discorso giuridico sia veicolata dalle specifiche costruzioni sintattiche rintracciate nei testi giuridici analizzati (Paragrafo 7.4);

- come l'assunto teorico proprio della 'Frame Semantics' (e realizzato in FrameNet) per cui una situazione conoscitiva può essere vista da più punti di vista, assumendo le diverse prospettive individuali delle entità coinvolte nella situazione–tipo generale, consenta di rintracciare nel testo la realizzazione dei diversi punti di vista prospettici riconducibili ad un unico concetto deontico (Paragrafo 7.5).

Sebbene la metodologia di annotazione semantica proposta in questo lavoro consista nel riutilizzare frames, FEs e STs definiti in FrameNet, tuttavia gli esperimenti di annotazione condotti hanno messo in luce la necessità di specializzare il modello originario. Per questo, una serie di proposte di specializzazioni della risorsa FrameNet sono esposte nel Paragrafo 7.6.

7.1 I frames 'deontici' in FrameNet

Fase preliminare del caso di studio è stata la verifica di quali tra i frames presenti in FrameNet consentano di descrivere le tre modalità di 'obbligo', 'permesso' e 'divieto'. I frames selezionati sono riportati rispettivamente nella Tabella 7.1, 7.2 e 7.3.

Modalità deontica: **obbligo**	
Frame	Definizione in FrameNet
OBLIGATION_SCENARIO (Non–Lexical Frame)	Under some, usually implicit, Condition a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable social Consequence for the Responsible_party. This Consequence may or may not be stated overtly.
BEING_OBLIGATED	Under some Condition, usually left implicit, a Responsible_party is required to perform some Duty. If they do not perform the Duty, there may be some undesirable Consequence, which may or may not be stated overtly.
BEING_OBLIGATORY	Under some Condition, usually left implicit, a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable Consequence for the Responsible_party, which may or may not be stated overtly. Compare this frame to the Being_obligated frame.
IMPOSING_OBLIGATION	A Duty is imposed on a Responsible_party according to a Principle which regulates how the Responsible_party should respond to a Situation. The Situation may be expressed metonymically by reference to an Obligator, whose action invokes the Principle. It is only rarely the case that the Principle and the Situation/Obligator are both expressed overtly.

Tabella 7.1: I frames in FrameNet che descrivono lo status di 'obbligo'.

Sono inoltre presenti in FrameNet altri due frames riconducibili a più di una modalità deontica (vedi Tabella 7.4). Nel caso del frame LAW, l'attenzione è posta sulla descrizione di una situazione–tipo nella quale una norma giuridica regola uno stato di cose che deve necessariamente essere tale (FE Required) o che è vietato (FE Forbidden). Nel frame LEGALITY, è descritto lo status di conformità o violazione di un'azione o di un oggetto rispetto alla norma giuridica che determina se essi siano permessi o vietati.

Modalità deontica: **permesso**	
PERMITTING	In this frame a State_of_affairs is permitted by a Principle. Raising constructions are common in this frame. In this frame the Principle which sanctions the State_of_affairs is not an agent who grants permission to a specific individual or group of individuals, and thus differs from the Grantor in the Grant_permission frame.

Tabella 7.2: Il frame in FrameNet che descrive lo status di 'permesso'.

Modalità deontica: **divieto**	
PROHIBITING	In this frame a State_of_affairs is prohibited by a Principle. Raising constructions are common in this frame. In this frame the Principle which prohibits the State_of_affairs is not an agent who denies permission to a specific individual or group of individuals, and thus differs from the Authority in the Deny_permission frame.
DENY_PERMISSION	In this frame, an Authority orders a Protagonist not to engage in an Action.

Tabella 7.3: I frames in FrameNet che descrivono lo status di 'divieto'.

Modalità deontica: **obbligo + divieto**	
LAW	A Law regulates activities or states of affairs within a Jurisdiction, dictating what Required states should be the case and what Forbidden states should not. Often it also indicates negative consequences for individuals that violate it, and these negative consequences are generally enforced by some official authority. They may or may not be created by some official legislative body.
Modalità deontica: **permesso + divieto**	
LEGALITY	Words in this frame describe the status of an Action with respect to a Code of laws or rules. An Object may also be in violation or compliance of the Code by virtue of its existence, location or possession.

Tabella 7.4: I frames in FrameNet riconducibili a più di una modalità deontica.

Per il fatto di essere legati da una relazione 'frame–to–frame' di tipo Using al frame OBLIGATION_SCENARIO, come si può vedere nella Figura 7.1, anche i frames riportati nella Tabella 7.5 sono qui ritenuti importanti per la piena

descrizione dello scenario d'**obbligo**.

COMPLIANCE	This frame concerns Acts and State_of_Affairs for which Protagonists are responsible and which either follow or violate some set of rules or Norms.
BEING_IN_EFFECT	A particular Binding_principle is (or is not) operative, that is, any obligations, restrictions, and any other aspects of the Binding_principle are (or are not) in effect. The Binding_principle can be expressed as being in effect for a particular Duration, or at a particular Time or Place, or under certain Circumstances.
DOCUMENTS	Words in the frame refer to any Document that has a legal status. Some Document empowers the Bearer of the Document to execute the Right. Others indicate the Obligation of the Bearer. Still others show the identity or Status of the Bearer.

Tabella 7.5: I frames legati da una relazione di tipo Using al frame OBLIGATION_SCENARIO.

Infine, sebbene il frame REQUIRED_EVENT (la cui definizione è riportata nella Tabella 7.6) non possa essere considerato rappresentativo di una situazione–tipo **deontica**, tuttavia si è ritenuto interessante annoverarlo tra i frames presi in esame in questo caso di studio, perché consente di descrivere la modalità **anankastica** di uno stato di cose. Come sarà infatti discusso oltre, è questa una dimensione complementare a quella deontica propria di un enunciato normativo. Non a caso infatti il frame è legato da una relazione di Inheritance al frame OBLIGATION_SCENARIO.

REQUIRED_EVENT	Unless a particular Required_situation obtains, Negative_consequences will follow. Alternatively, the Required_situation is required to achieve a Purpose (which avoids Negative_consequences). A set of Circumstances may be specified under which the requirement holds.

Tabella 7.6: Il frame 'anankastico' REQUIRED_EVENT.

7.1.1 Le relazioni 'frame–to–frame'

I frames qui considerati sono legati dalle relazioni 'frame–to–frame' elencate nella Tabella 7.7[3].

Super Frame	**Relazione**	**Sub Frame**
Required_event	Is_inherited_by	Obligation_scenario
Obligation_scenario	Is_perspectivized_in	Being_obligatory, Being_obligated
Imposing_obligation	Is_causative_of	Being_obligated
Obligation_scenario	Is_used_by	Compliance, Documents, Being_in_effect
Being_in_effect	Is_used_by	Enforcing
Compliance	Is_used_by	Legality
Law	Is_used_by	Legality, Prohibiting
Prohibiting	Is_inherited_by	Permitting
Communication	Is_used_by	Deny_permission, Grant_permission

Tabella 7.7: I tipi di relazioni 'frame–to–frame' presenti in FrameNet tra i frames considerati 'deontici'.

Come si può vedere nella Figura 7.1, realizzata grazie al FrameGrapher, tali relazioni permettono di rappresentare in modo formale la conoscenza deontica in una rete organizzata di frames.

7.2 Il punto di partenza: l'annotazione semantica della struttura sintattica a dipendenze

Il punto di partenza della metodologia di annotazione semantica messa a punto in questo lavoro è rappresentato dall'output dell'annotazione sintattica a dipendenze realizzata da DeSR[4]. Di fatto, l'informazione relativa ai frames presenti in un periodo è aggiunta in modo manuale sulla struttura sintattica ad albero generata in modo automatico dal parser.

[3]La descrizione dei singoli tipi di relazione è esposta nel Paragrafo 7.5 e seguenti.
[4]Le descrizioni che seguono riprendono parti di quelle proposte da Venturi (2011).

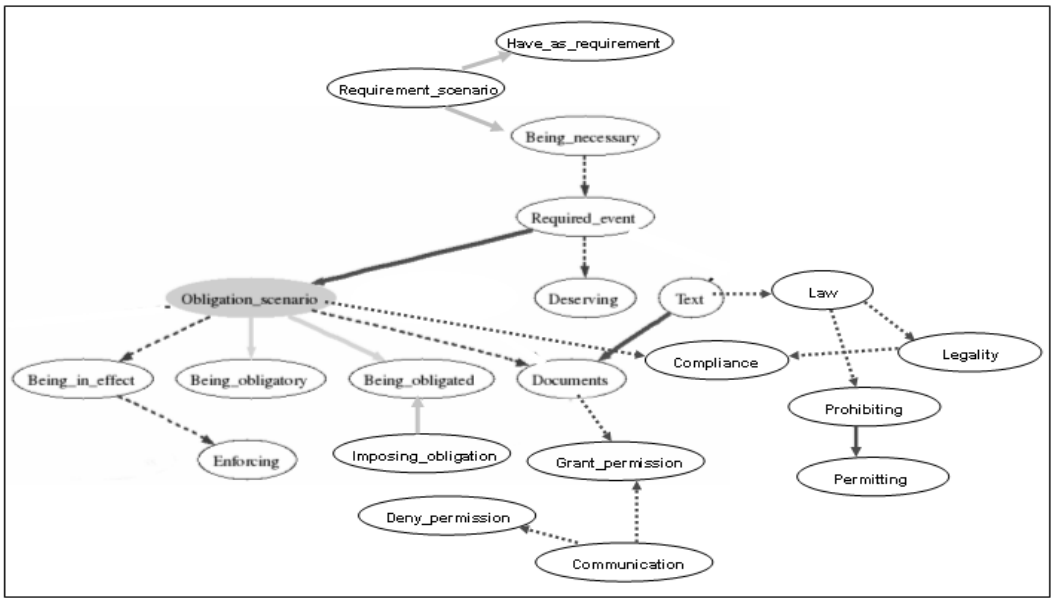


Figura 7.1: La rete di relazioni 'frame–to–frame' che lega i frames 'deontici' in FrameNet.

Ne è un esempio l'annotazione del seguente periodo riportata nella Figura 7.2:

a) *Obbligati al pagamento della tassa sono gli esercenti i grandi impianti di combustione di cui all'articolo 1.*

I criteri adottati in fase di annotazione sono i seguenti:

- ogni frame è annotato a partire da un singolo token (la LU evocatrice) da cui dipende un albero o un sotto–albero sintattico. In questo caso, il frame BEING_OBLIGATED è evocato dal participio passato *obbligati*, radice (root) dell'intero periodo, e il frame COMMERCE_PAY è evocato dal sostantivo *pagamento* dal quale dipende il sotto–albero sintattico *della tassa* tramite una relazione di tipo 'comp';

- ogni FE è annotato a partire da una determinata relazione di dipendenza che lega una porzione di testo al token evocatore. Questo permette

di mettere in collegamento i ruoli semantici svolti in un determinato frame con le corrispondenti relazioni di dipendenza.

Ad esempio, la porzione di testo *gli esercenti i grandi impianti di combustione di cui all'articolo 1*, legata alla LU *obbligati* da una relazione di dipendenza di tipo 'subj_pass', rappresenta un'istanza del FE Responsible_party del frame BEING_OBLIGATED. E la porzione di testo *della tassa*, legata alla LU *pagamento* da una relazione di tipo 'comp', ricopre il ruolo semantico (FE) Money previsto nel frame COMMERCE_PAY;

- i casi di 'Null Instantiation' di FEs sono trattati come casi sottospecificati. Un esempio è rappresentato dal FE Buyer parte del frame COMMERCE_PAY. In questo caso il *compratore*, agente della situazione, non è lessicalmente espresso ma è chiaramente desumibile dalla struttura sintattica (a soggetto controllato) dell'intero periodo. Si tratta infatti di un esempio di 'Constructional Null Instantiation' (CNI), annotato in modo sottospecificato sulla LU evocatrice *pagamento*.

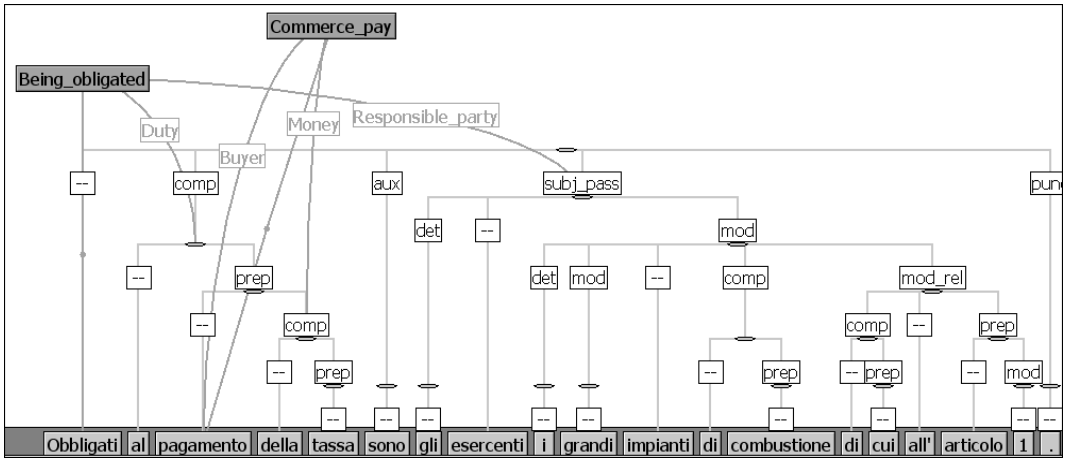


Figura 7.2: Un esempio di periodo annotato a livello sintattico e semantico.

Nel caso in cui la LU evocatrice sia rappresentata da un'unità lessicale polirematica, costituita cioè da una costruzione a verbo supporto, i criteri di annotazione sono stati parzialmente modificati. Ne è un esempio l'annotazione del seguente periodo riportata nella Figura 7.3:

b) *Gli enti gestori delle reti ed opere destinate al pubblico servizio in qualsiasi modo interferenti con l'infrastruttura da realizzare hanno l'obbligo di cooperare alla realizzazione della stessa con le modalità previste dall'articolo 5, come precisato dal presente articolo.*

In questo caso il frame BEING_OBLIGATED è evocato dal sostantivo *obbligo* unito al verbo supporto *avere* da una relazione di dipendenza 'obj' rispetto alla quale, sulla base dei criteri di annotazione sintattica, il verbo costituisce la testa sintattica e il sostantivo il dipendente. Di conseguenza, il FE Duty è istanziato dalla porzione di testo *di cooperare alla realizzazione della stessa con le modalità previste dall'articolo 5*, dipendente dal token *obbligo* con una relazione di dipendenza di tipo 'arg'; il FE Responsible_party è istanziato dal soggetto del verbo supporto, cioè dalla porzione di testo *gli enti gestori delle reti ed opere destinate al pubblico servizio in qualsiasi modo interferenti con l'infrastruttura da realizzare* dipendente dal token *hanno* tramite la relazione 'subj'.

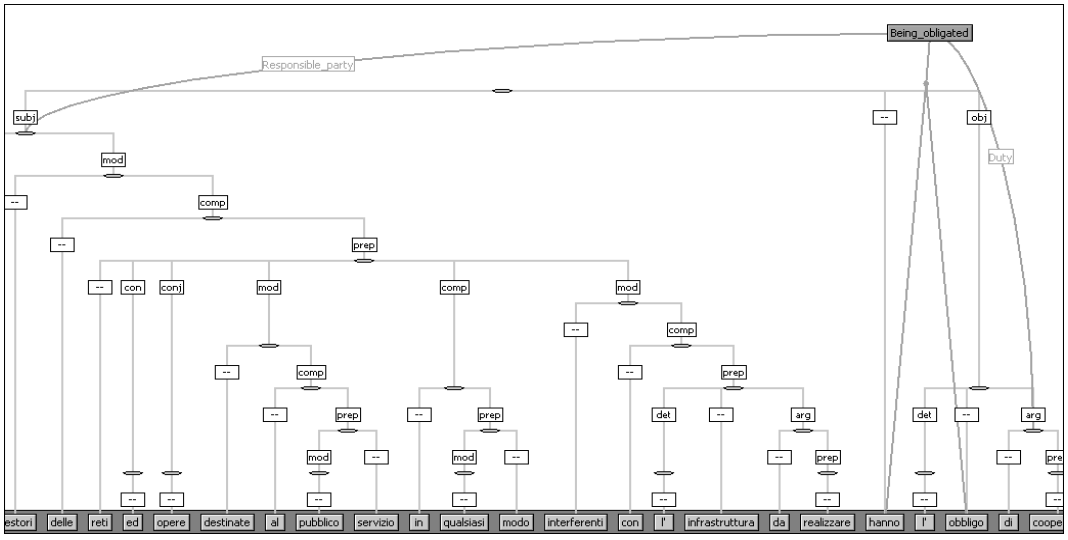


Figura 7.3: Un esempio di periodo annotato a livello sintattico e semantico con LU *(avere)obbligo*.

Dunque, a differenza dell'esempio a), in b) i FEs non stati annotati unicamente su dipendenti della LU monorematica evocatrice del frame BEING_OBLIGATED,

ma a partire da entrambi gli elementi che fanno parte della LU polirematica. È in questo modo rispettata la struttura sintattica del periodo e, nello stesso tempo, ne sono resi espliciti tutti gli elementi che contribuiscono alla rappresentazione del suo contenuto informativo.

L'intera strategia di annotazione è espressamente finalizzata a superare l'ostacolo messo in evidenza da Dolbey (2009) riguardo al fatto che "FrameNet annotations are not linked to syntactic parse trees" e che di conseguenza spesso le istanze dei FEs non corrispondono "to syntactic constituents provided by a syntactic parse of the whole sentence". Come precedentemente discusso nel Paragrafo 6.3.4, l'assenza di informazione sulla struttura sintattica globale della frase può creare alcune difficoltà nello svolgimento di compiti di gestione dell'informazione basati sull'annotazione linguistica del testo.

Un tale approccio riecheggia *i)* quello messo a punto da Burchardt et al. (2009) nell'ambito del progetto SALSA, finalizzato all'annotazione manuale di un corpus di articoli giornalistici in lingua tedesca con informazione 'a frame', corpus precedentemente annotato a livello sintattico e *ii)* quello seguito dal gruppo di ricerca dell'Università di Pisa attivo nell'ambito del progetto IFrame e attualmente impegnato nell'annotazione semantica 'a testo continuo' di una porzione della "Italian Syntactic–Semantic Treebank" (ISST).

La strategia di annotazione messa a punto in questo lavoro si differenzia, tuttavia, da questi due casi per un aspetto fondamentale. L'annotazione semantica si basa sul risultato della fase di analisi sintattica automatica realizzata dal parser e non su un corpus preesistente annotato sintatticamente (su di una treebank), come stabilito invece nei progetti SALSA e IFrame. Come discusso nel Paragrafo 3.3, ciò è dovuto all'attuale mancanza di una treebank del tipo di testi normativi al centro dell'esperimento di annotazione semantica realizzato in questo lavoro.

Tenuto in considerazione l'impatto che le caratteristiche della lingua del diritto hanno sulle performances di DeSR, l'annotazione semantica è stata preceduta da una fase di revisione manuale degli errori commessi dal parser in fase di annotazione sintattica automatica.

Infine, come molti dei progetti descritti nel Paragrafo 6.1.3 finalizzati all'uso del modello FrameNet, anche in questo lavoro è stato utilizzato lo strumento di annotazione e visualizzazione grafica SALSA tool (Erk et al., 2003), sviluppato nell'ambito del progetto SALSA. Gli esempi di annotazione riportati nelle Figure 7.2 e 7.3 sono stati infatti realizzati con il SALSA tool.

7.3 Le modalità di annotazione

Sulla base di questi criteri di annotazione, sono stati condotti tre diversi esperimenti di annotazione del corpus AMBnorm(Stato). Pur non volendo essere esaustivi, tali esperimenti hanno l'obiettivo di illustrare come l'annotazione semantica basata sul modello FrameNet consenta di accedere al contenuto dei testi in esame a diversi livelli: *i)* a livello dell'informazione lessicografica veicolata da alcuni dei termini più significativi in essi contenuti (grazie alla modalità di annotazione lessicografica), *ii)* a livello della rappresentazione esplicita di quali sono in un periodo tutti gli elementi lessicali che contribuiscono alla completa comprensione del suo contenuto informativo (grazie all'annotazione 'a testo continuo').

Un caso particolare di *ii)* è poi quello affrontato nel Paragrafo 7.3.3, dove sono riportati alcuni esempi di annotazione condotti sulla base di una modalità innovativa. Tale modalità è stata messa a punto in questo studio con il preciso obiettivo di cercare di rendere esplicito l'intrecciarsi di realtà espressione del mondo del diritto e del mondo regolato dal diritto, contribuendo in questo modo alla completa rappresentazione del contenuto informativo dei periodi annotati.

I tre esperimenti hanno portato all'annotazione di 124 periodi, per un totale di 65 tipi diversi di frames annotati e 192 di istanze lessicali (LUs) diverse. Mentre l'elenco completo dei frames annotati e delle relative LUs è riportato nell'Allegato II, in quanto segue sono discussi alcuni esempi di annotazione.

7.3.1 L'annotazione lessicografica

L'adozione di questa modalità di annotazione ha due obiettivi. Da un lato, è finalizzata ad illustrare in che modo la selezione a priori di LUs evocatrici dei frames 'deontici' selezionati abbia guidato l'intero caso di studio, contribuendo a rendere esplicito come l'informazione deontica contenuta in AMBnorm(Stato) sia lessicalmente istanziata (Paragrafo 7.3.1.1).

Dall'altro, ha la finalità di mostrare come una tale modalità di annotazione renda possibile acquisire da un testo il significato di una parola (cioè, di una LU) raccogliendone le proprietà semantico–sintattico combinatorie (Paragrafo 7.3.1.2).

7.3.1.1 La selezione delle LUs evocatrici

Il processo di selezione delle LUs evocatrici di frames deontici è stato guidato, oltre che dall'intuizione linguistica, anche dalla consultazione della lista di LUs previste in FrameNet per i frames scelti, dalle parole parte dei synsets di JurWordNet, nonché dalla consultazione di dizionari della lingua italiana e dal "Dizionario giuridico" (Edizione Simone) nella sua versione on–line [5].

Inoltre, come segnalato nella lista riportata nell'Allegato II, alcune delle LUs poliremathiche sono state selezionate in modo automatico, estraendole da AMBnorm(Stato). A questo scopo è stata messa a punto una metodologia di selezione semi–automatica, che, utilizzando l'approccio all'estrazione automatica di terminologia descritto nel Paragrafo 5.2.1, ha permesso di acquisire in modo automatico dal testo annotato a livello morfosintattico sequenze di categorie morfosintattiche del tipo verbo/sostantivo, preposizione/sostantivo, ecc...

Modificando i filtri linguistici usati, sono state considerate rilevanti le sequenze qui di seguito elencate:

- verbo+sostantivo, es. *fare obbligo, avere efficacia*;
- verbo+articolo+sostantivo, es. *assumere l'obbligo, definire l'obbligo*;
- verbo+sostantivo+preposizione, es. *fare luogo a, essere soggetto a*;
- verbo+preposizione+sostantivo, es. *entrare in vigore*;
- preposizione+sostantivo, es. *in violazione, in vigore.*

Come si può notare dagli esempi riportati, sono questi tutti casi in cui l'informazione semantica è veicolata non dalla testa sintattica dell'unità lessicale, cioè non dal verbo o dalla preposizione, ma dal sostantivo.

L'obiettivo che ha guidato la definizione di questa metodologia di selezione di LUs era quello di dimostrare come alcune delle caratteristiche morfosintattiche messe in luce in fase di monitoraggio dei corpora giuridici avessero delle conseguenze rilevanti a livello di organizzazione del materiale semantico–lessicale nel testo. La bassa percentuale di occorrenza di verbi e l'elevata presenza di sostantivi e preposizioni sono state considerate le spie principali. L'idea inoltre era quella di verificare quanto affermato da Rovere

[5] http://www.simone.it/newdiz/newdiz.php?action=view&id=1&title=A%20non%20domino

(2005, p. 158 e sgg.) a proposito del "depotenziamento semantico" di alcuni verbi nei testi giuridici, di quei verbi cioè che, utilizzati in costrutti tecnici per lo più con "finalità sintattiche", perdono il loro valore semantico originario.

Ne è stata una conferma la possibilità di riuscire ad individuare sequenze di verbi, sostantivi e preposizioni che, co–occorrendo in AMBnorm(Stato) con valori di forza di associazione statisticamente significativi, costituiscono unità terminologiche dotate di una struttura predicativa e in grado di evocare una situazione–tipo[6]. La riprova di quanto osservato da Rovere (2005) sta in particolare nel fatto che nel caso delle costruzioni a verbo supporto identificate, il verbo perde il proprio potere semantico e la capacità evocativa passa al sostantivo.

Una tale metodologia è diretta conseguenza del carattere 'formulaico' della lingua del diritto di cui AMBnorm(Stato) è qui rappresentativo. Come suggerito da Eklund-Braconi (2000), il contenuto semantico–lessicale di una frase di un testo giuridico è infatti veicolato da sequenze di singole unità lessicali che costituiscono vere e proprie "formule", più o meno fisse, dotate di un "significato finito e specialistico". La strategia di mettere in relazione il calcolo della forza associativa dei singoli elementi costitutivi di tali "formule" e la loro capacità di evocare una situazione conoscitiva ne è dunque una conferma.

Essa è inoltre in linea con la concezione abbracciata nel progetto FrameNet per cui ogni unità mono e polirematica, nonché ogni espressione idiomatica, è da considerarsi una 'parola' dotata di una propria struttura predicativa[7]. Ciò dimostra come i principi organizzativi di FrameNet lo rendano un modello particolarmente espressivo per rendere esplicito in un testo giuridico il legame tra strutture linguistiche e significato.

Tale approccio alla selezione di LUs polirematiche ha permesso, inoltre, di rendere esplicito come verbi supporto diversi siano dirimenti per evocare situazioni conoscitive diverse. È il caso, ad esempio, del sostantivo *obbligo*, che quando oggetto del verbo supporto *avere* rimanda al frame BEING_OBLIGATED, mettendo a fuoco un particolare punto di vista sullo scenario deontico generale; mentre, quando in unione al verbo supporto *definire* evoca il frame IMPOSING_OBLIGATION, veicolando una diversa prospettiva di osservazione.

[6]La verifica della capacità evocatrice di tali unità polirematiche è stata manuale.
[7]Vedi Ruppenhofer et al. (2010, pp. 7–8).

7.3.1.2 Un esempio di entrata lessicografica

In questo paragrafo è fornito un esempio di come si presenterebbe l'entrata lessicale del sostantivo *obbligo* in un lessico giuridico costruito sui principi di organizzazione del significato di FrameNet. Ci si è in particolare concentrati sui casi in cui *obbligo* evoca i frames BEING_OBLIGATED e IMPOSING_OBLIGATION. A partire dai periodi annotati in AMBnorm(Stato), sono stati dunque raccolti i vari tipi di relazione di dipendenza legati alla testa sintattica *obbligo*, espressione sintattica dei FEs dei due frames[8].

I risultati sono riportati rispettivamente nelle Tabelle 7.8 e 7.9[9] e sono da accompagnare ai seguenti esempi di annotazione[10]:

a) *Le attività di trasporto e dispacciamento del gas naturale a rete, nonché la gestione di infrastrutture di approvvigionamento di energia connesse alle attività di trasporto e dispacciamento di energia a rete, sono di interesse pubblico e* [*sono sottoposte* $_{Supp}$][11] *agli* ***obblighi*** [*di servizio pubblico* $_{Duty}$] *derivanti dalla normativa comunitaria, dalla legislazione vigente e da apposite convenzioni con le autorità competenti.*

b) *Gli enti gestori delle reti ed opere destinate al pubblico servizio in qualsiasi modo interferenti con l'infrastruttura da realizzare* [*hanno* $_{Supp}$] *l'****obbligo*** [*di cooperare alla realizzazione della stessa con le modalità previste dall'articolo 5* $_{Duty}$], *come precisato dal presente articolo.*

c) [*Le figure soggettive esercenti pubblici servizi o titolari di pubbliche funzioni* $_{Responsible\_party}$] [*hanno* $_{Supp}$] *l'****obbligo****, sulla base di accordi commerciali a condizioni eque e non discriminatorie, di consentire l'accesso*

[8]Nota che, come spiegato nel Paragrafo 7.2, *obbligo* non è testa sintattica dei FEs ad esso legati nel caso in cui la LU sia inserita in una costruzione supporto. È il caso, ad esempio, dell'esempio b), nel quale il FE Responsible_party (*Gli enti gestori delle reti ed opere destinate al pubblico servizio in qualsiasi modo interferenti con l'infrastruttura da realizzare*) è legato da una relazione di dipendenza 'subj' al verbo supporto *hanno* e non a *obbligo*.

[9]In grassetto sono segnalati i 'Core' FEs.

[10]Come per i successivi esempi di annotazione, è stata qui adottata la convenzione usata in FrameNet di segnalare la LU evocatrice in grassetto, di racchiudere la realizzazione testuale dei FEs tra parentesi quadre e di riportare il nome del FE a pedice.

[11]Nota che in questo modo in FrameNet viene segnalata la presenza di materiale lessicale supporto della LU evocatrice. Per coerenza, la stessa notazione è stata adottata in questo studio.

alle proprie infrastrutture civili disponibili, a condizione che non venga turbato l'esercizio delle rispettive attività istituzionali.

d) *L'**obbligo** di versamento non si applica agli impianti o alle infrastrutture per i quali alla data di entrata in vigore della presente legge si sia già conclusa l'istruttoria.* [DNI $_{Responsible\_party}$]

e) *Il contraente generale* [*assume* $_{Supp}$] *l'**obbligo** di verificare il progetto esecutivo posto in gara e di farlo proprio,* [*fermo restando quanto disposto dal comma 5 dell'articolo 9* $_{Condition}$].

f) [*In caso di mancato rispetto del programma di cui al comma 4, ovvero di mancata segnalazione ai sensi del comma 2,* $_{Condition}$] *il soggetto gestore* [*ha* $_{Supp}$] *l'**obbligo** di risarcire i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori.*

g) [*È fatto* $_{Supp}$] ***obbligo** a chiunque spetti di osservarlo e di farlo osservare.* [DNI $_{Obligator}$]

h) [*È fatto* $_{Supp}$] ***obbligo*** [*ai rivenditori dei beni di cui al comma 2* $_{Responsible\_party}$] *di accettare la restituzione di analogo bene usato, purché presente nel loro assortimento, anche se di marca o tipo diversi.*

i) *In conformità alla vigente normativa in materia di smaltimento dei rifiuti,* [*è fatto* $_{Supp}$] ***obbligo** a tutti i detentori di prodotti, di impianti e di beni durevoli contenenti le sostanze lesive* [*di conferire i medesimi, al termine della loro durata operativa, a centri di raccolta autorizzati* $_{Duty}$].

l) [*Nel caso di affidamento dei lavori in assicurazione di qualità,* $_{Condition}$][12] [*qualora la stazione appaltante non abbia già adottato un proprio sistema di qualità,* $_{Condition}$][13] [*è fatto* $_{Supp}$] ***obbligo** alla stessa di affidare, ad idonei soggetti qualificati, secondo le procedure di cui al decreto legislativo 17 marzo 1995, n. 157, i servizi di supporto al responsabile del procedimento ed al direttore dei lavori, in modo da assicurare che anche il funzionamento della stazione appaltante sia conforme ai livelli di qualità richiesti dall'appaltatore.*

[12]Realizzato come 'comp'.
[13]Realizzato come 'arg'.

m) [*È fatto* $_{Supp}$] ***obbligo*** *ai comuni di adeguare gli strumenti urbanistici ai fini di rendere possibile lo scorporo dal calcolo della superficie utile e del volume edificato degli spessori di chiusure opache verticali ed orizzontali nei limiti più avanti precisati,* [*al fine di favorire la realizzazione di edifici con adeguata inerzia termica e sfasamento termico* $_{Purpose}$].

FE	Realizzazione sintattica	No. di istanze	Esempio
Duty	comp	5	a)
	arg	8	b)
Responsible_party	subj	13	c)
	DNI	2	d)
Condition	mod	2	e)
	comp	4	f)

Tabella 7.8: Realizzazione sintattica dei FEs legati a *obbligo* nei periodi in cui esso evoca il frame BEING_OBLIGATED.

FE	Realizzazione sintattica	No. di istanze	Esempio
Obligator	CNI	6	g)
Responsible_party	comp	6	h)
Duty	arg	5	i)
Condition	comp	2	l)
	mod	1	l)
Purpose	comp	1	m)

Tabella 7.9: Realizzazione sintattica dei FEs legati a *obbligo* nei periodi in cui esso evoca il frame IMPOSING_OBLIGATION.

La raccolta di questi dati permette di mettere in luce come a seconda del frame evocato, il termine *obbligo* assuma comportamenti sintattici diversi. Così, ad esempio, in un contesto nel quale esso rimanda alla situazione–tipo descritta dal frame BEING_OBLIGATED il FE Responsible_party rappresenta il soggetto ('subj') della costruzione a verbo supporto nel quale *obbligo* è inserito, ma può anche essere omesso. In questo caso, un tale comportamento viene comunque annoverato tra quelli possibili ed è espresso come un caso di 'Definite Null Instantiation' (DNI), desumibile dal contesto in maniera anaforica.

Il fatto di poter rendere conto anche di quest'ultimo tipo di comportamento è di estrema importanza in un'ottica di futura costruzione di un lessico giuridico basato sui principi di organizzazione del significato messi a punto nel progetto FrameNet. È questo infatti un tipo di informazione che raramente viene considerata nei dizionari tradizionali e che al contrario FrameNet permette di considerare tra le possibili proprietà combinatorie di un lemma[14].

Estendendo questo tipo di raccolta dati a tutte le LUs più significative, la finalità di questo tipo di annotazione è duplice. Da un lato, essa, ponendosi l'obiettivo di estrarre da un corpus tutte le proprietà combinatorie di una parola, permette di metterne in luce eventuali comportamenti sintattici idiosincratici legati al contenuto informativo del periodo. Dall'altro, pone le basi per la definizione di una metodologia di annotazione automatica di ruoli semantici ('Automatic Semantic Role Labeling'), basandola sul livello di annotazione sintattica a dipendenze.

7.3.2 L'annotazione 'a testo continuo'

Nel secondo esperimento di annotazione condotto, è stata adottata una modalità al centro delle più recenti attività del gruppo di Berkeley: l'annotazione 'a testo continuo'. Come precedentemente ricordato[15], essa è stata messa a punto nell'ambito del progetto FrameNet con l'esplicito intento di dimostrare come i principi di organizzazione del significato propri della 'Frame Semantics Theory' forniscano uno strumento affidabile di rappresentazione dell'intero contenuto di un periodo.

Una tale modalità è stata dunque sperimentata in questo lavoro con l'obiettivo di verificare come essa possa essere applicata con successo nel caso di testi giuridici.

Come mostra l'esempio che segue, l'annotazione 'a testo continuo' permette di rendere esplicito l'intero contenuto informativo nel periodo a). Per chiarezza espositiva, sono stati prima segnalate nel periodo tutte le LUs considerate evocatrici (in grassetto), seguite dal corrispondente frame (tra paren-

[14]Fanno notare a questo proposito Atkins et al. (2003b): "Though an extremely common phenomenon, null instantiation has left few traces in dictionaries: if it is handled at all, it is certainly not dealt with any systematic way. [...] FrameNet's contribution here is to draw our attention to the significance of null instantiation to a succesful description of many classes of words.".

[15]Vedi il Paragrafo 6.1.2.

tesi quadre). In a.1), a.2), a.3) e a.4) sono poi state riportate le annotazione relative ai singoli frames.

a) *Con il* ***provvedimento***[LAW] *di sospensione la sezione regionale* ***assegna***[IMPOSING_OBLIGATION] *un termine, che non può comunque superare i dodici mesi, entro il quale l'impresa o l'ente iscritto* ***deve***[BEING_OBLIGATED] ***conformare***[COMPLIANCE] *alla normativa vigente l'attività ed i suoi effetti.*

   a.1) [LAW] [***provvedimento*** $_{Law}$][16] [*di sospensione* $_{Required}$]

   a.2) [IMPOSING_OBLIGATION] [*Con il provvedimento di sospensione* $_{Means}$] [*la sezione regionale* $_{Obligator}$] ***assegna*** [*un termine, che non può comunque superare i dodici mesi, entro il quale l'impresa o l'ente iscritto deve conformare alla normativa vigente l'attività ed i suoi effetti* $_{Duty}$]. [CNI $_{Responsible\_party}$]

   a.3) [BEING_OBLIGATED] [*entro il quale* $_{Time}$] [*l'impresa o l'ente iscritto* $_{Responsible\_party}$] ***deve*** [*conformare alla normativa vigente l'attività ed i suoi effetti* $_{Duty}$]

   a.4) [COMPLIANCE] ***conformare*** [*alla normativa vigente* $_{Norm}$] [*l'attività ed i suoi effetti* $_{State\_of\_affairs}$] [CNI $_{Protagonist}$][17]

In questo modo è stato possibile esplicitare il contributo che ogni unità lessicale predicativa porta alla comprensione del contenuto proposizionale del periodo a). Sono così resi espliciti tutti gli elementi di conoscenza che contribuiscono alla rappresentazione esaustiva di tutte le situazioni–tipo necessarie per comprendere che ci si sta riferendo al fatto che una pubblica autorità (*la sezione regionale*: FE Obligator in IMPOSING_OBLIGATION), mediante un atto giuridico (il *provvedimento*: FE Law in LAW, FE Means in IMPOSING_OBLIGATION), esplica il proprio potere su un soggetto imponendo un obbligo (il *conformare alla normativa vigente l'attività ed i suoi effetti*: FE Duty in IMPOSING_OBLIGATION). Ciò implica che il soggetto (*l'impresa o l'ente iscritto*: FE Responsible_party in BEING_OBLIGATED) è obbligato a svolgere il dovere imposto (FE Duty in BEING_OBLIGATED), che nella fattispecie è quello di conformare un insieme di azioni (*l'attività ed i suoi effetti*:

[16] In questo caso la LU coincide con l'istanza di un FE del frame.

[17] Si tratta di un caso di 'Constructional Null Instantiation' legata alla struttura sintattica a soggetto controllato.

FE State_of_affairs in COMPLIANCE) ad un principio normativo (la *normativa vigente*: FE Norm in COMPLIANCE) entro un determinato termine di tempo.

Inoltre, una tale modalità di annotazione è adottata nell'ambito delle attività del progetto FrameNet finalizzate a dimostrare come i principi di annotazione semantica messi a punto possano essere usati per svolgere un vero e proprio compito di comprensione testuale, che travalichi dunque i confini del singolo periodo[18]. Per questo motivo, essa è stata qui sperimentata per verificare come possa essere adattata all'annotazione semantica di quelle parti di testo che, sulla base della scelta di segmentazione del testo in periodi descritta nel Paragrafo 3.3.1.1, sono state suddivise in più periodi distinti.

Un esempio è rappresentato dal seguente periodo:

b) *Le società e gli enti gestori di servizi pubblici di trasporto o delle relative infrastrutture, inclusi i comuni, le province e le regioni, hanno l'obbligo di:*

- *individuare le aree in cui per effetto delle immissioni delle infrastrutture stesse si abbia superamento dei limiti di immissione previsti;*
- *determinare il contributo specifico delle infrastrutture al superamento dei limiti suddetti.*

Esso, sulla base dei criteri imposti, è stato segmentato in tre periodi diversi, che sono stati semanticamente annotati come segue[19]:

b.1) [BEING_OBLIGATED] [*Le società e gli enti gestori di servizi pubblici di trasporto o delle relative infrastrutture, inclusi i comuni, le province e le regioni,* $_{Responsible\_party}$] [*hanno* $_{Supp}$] *l'**obbligo** di:* [DNI $_{Duty}$]

b.2) [LOCATING] - ***individuare*** [*le aree in cui per effetto delle immissioni delle infrastrutture stesse si abbia superamento dei limiti di immissione previsti* $_{Sought\_entity}$]; [CNI $_{Perceiver}$]

b.3) [DECIDING] - ***determinare*** [*il contributo specifico delle infrastrutture al superamento dei limiti suddetti* $_{Decision}$]. [CNI $_{Cognizer}$]

[18] Vedi Paragrafo 6.1.3.

[19] Nota che, come era stato fatto osservare nel Paragrafo 3.3.1.1, la scelta di mantenere l'originaria segmentazione del testo in più periodi distinti era stata mossa dall'esplicito intento di preservare l'originaria organizzazione dell'informazione voluta dal legislatore.

Assumendo dunque un'ottica di comprensione testuale che vada oltre il singolo periodo, un'annotazione di questo tipo apre la strada ad un processo di successiva ricostruzione dei rapporti tra frames annotati in periodi diversi. Ciò permetterebbe infatti di rendere esplicito il fatto che l'obbligo a cui si fa riferimento in b.1), obbligo il quale un soggetto giuridico (*le società e gli enti gestori di servizi pubblici di trasporto o delle relative infrastrutture, inclusi i comuni, le province e le regioni*: FE Responsible_party in BEING_OBLIGATED) è tenuto ad adempiere, è rappresentato dalle situazioni–tipo descritte in b.2) e b.3). In particolare, l'annotazione rende esplicito che il dovere (FE Duty in BEING_OBLIGATED) non è lessicalmente istanziato nel periodo, ma è comunque desumibile in maniera anaforica dal contesto, più ampio di quello del singolo periodo.

Nella fattispecie, i soggetti obbligati (FEs Perceiver in LOCATING e Cognizer in DECIDING)[20] sono tenuti a *i)* localizzare con successo *le aree in cui per effetto delle immissioni delle infrastrutture stesse si abbia superamento dei limiti di immissione previsti* (FE Sought_entity in LOCATING) e *ii)* stabilire quale sia il valore indicativo del *contributo specifico delle infrastrutture al superamento dei limiti suddetti* (FE Decision in DECIDING).

7.3.3 L'annotazione di conoscenza 'giuridica' e 'extragiuridica'

Il terzo esperimento di annotazione condotto in questo caso di studio è consistito nell'annotare i frames evocati da LUs contenute in istanze di uno stato di cose 'obbligato', 'permesso' o 'vietato'. Ciò ha permesso di analizzare come la semantica delle situazioni 'del mondo' regolate si intrecci con quella delle regole di comportamento.

Una tale modalità di annotazione è nuova rispetto alle due messe a punto in FrameNet. Essa è stata ispirata dalla questione discussa nel Paragrafo 5.1.1 circa la compresenza in ogni enunciato normativo di una "componente semantica referenziale e una componente deontica", riflesso del "complesso intreccio di realtà giuridiche ed extragiuridiche" (Belvedere, 1994a) caratteristico di ogni discorso giuridico. Come precedentemente messo in luce si tratta di una questione aperta e ampiamente dibattuta in materia di rap-

[20] In questo caso la natura di 'Constructional Null Instantiation' delle istanze di questi due FEs è data dalla costruzione a verbo supporto nella quale occorrono, ricostruibile in questa fase di (ri)composizione del contenuto dell'intera sezione di testo.

presentazione formale della conoscenza giuridica alla quale si cerca qui di suggerire una possibile soluzione.

In base alla metodologia messa a punto, l'annotazione si è svolta in due fasi consecutive. Dopo una prima fase di annotazione di tipo lessicografico condotta a partire dalle LUs precedentemente selezionate in quanto evocatrici di realtà 'giuridiche' (deontiche), è stata resa esplicita l'informazione relativa alle realtà extragiuridiche regolate.

È il caso, ad esempio, del seguente periodo che in una prima fase è stato annotato come segue, rendendo esplicita la situazione deontica evocata dal participio passato *obbligato*:

a) [BEING_OBLIGATED] [*Qualora, in attuazione delle disposizioni del comma 2, siano avviate al consumo in rete miscele combustibile diesel–biodiesel con contenuto in biodiesel in misura superiore al 5 per cento* $_{Condition}$], [*i punti vendita nei quali tali miscele sono distribuite* $_{Responsible\_party}$] *sono* ***obbligati*** [*ad esporre idonee etichette di descrizione del prodotto, unitamente all'elenco dei veicoli omologati per l'uso dei predetti biocarburanti* $_{Duty}$].

In una seconda fase, è stata poi annotata l'informazione relativa al 'dovere' che *i punti vendita* sono obbligati ad adempiere. Pertanto, a partire dall'istanza del FE Duty nella frase è stata rintracciata la LU evocatrice (il verbo *esporre*) della situazione regolata (espressa dal frame CAUSE_TO_PERCEIVE) e al periodo a) è stata aggiunta la seconda seguente annotazione:

a.1) [CAUSE_TO_PERCEIVE] *Qualora, in attuazione delle disposizioni del comma 2, siano avviate al consumo in rete miscele combustibile diesel–biodiesel con contenuto in biodiesel in misura superiore al 5 per cento, i punti vendita nei quali tali miscele sono distribuite sono obbligati ad* ***esporre*** [*idonee etichette di descrizione del prodotto, unitamente all'elenco dei veicoli omologati per l'uso dei predetti biocarburanti* $_{Phenomenon}$]. [CNI $_{Actor}$]

Come ci si poteva aspettare, sulla base di questa metodologia di annotazione gli elementi che in ogni frame svolgono il ruolo tematico astratto di 'agenti' sono realizzati come frasi a soggetto controllato. Come esemplificato dal periodo a), il soggetto sintattico del verbo *esporre* è infatti soggetto del verbo deontico *obbligare*. Sebbene dunque il soggetto sia di fatto omesso in

a.1), ciononostante il suo contributo semantico (come FE Actor) alla descrizione del frame CAUSE_TO_PERCEIVE è reso esplicito grazie all'annotazione di un'istanza di 'Constructional Null Instantiation' (CNI).

In particolare, questo tipo di annotazione ha permesso di mettere in luce che i comportamenti imposti, permessi o vietati sono di tipo diverso. Sulla base delle opposizioni proposte da Belvedere (1994a) tra realtà giuridica ed extragiuridica a cui i termini del lessico giuridico fanno riferimento, 'doveri', 'permessi' e 'divieti' annotati sono stati classificati in:

- 'giuridici', quando evocati da termini usati per riferirsi a situazioni sempre parte del mondo del diritto;
- 'fattuali generici', quando evocati da termini che fanno riferimento a situazioni 'extragiuridiche' generali;
- 'fattuali specialistici', quando evocati da termini che fanno riferimento a situazioni 'extragiuridiche' specifiche del mondo di fatti reali regolati. Nel caso del corpus AMBnorm(Stato), sono per lo più situazioni legate alla materia ambientale legislata.

Nei paragrafi che seguono sono riportati esempi di annotazione delle tre tipologie di 'doveri, permessi, divieti' individuate.

7.3.3.1 L'annotazione di 'doveri'

Gli esempi di annotazione qui riportati dimostrano come i diversi tipi di 'doveri' imposti facciano riferimento a realtà riconducibili alle seguenti situazioni–tipo:

- situazioni descrittive di doveri 'giuridici', istanze di frames relativi a pratiche del mondo del diritto, quali l'adozione di regole di condotta (periodo a)), l'imposizione di una ammenda pecuniaria (periodo b)) o l'osservanza di regole di condotta (periodo c)):

  a) [ADOPT_SELECTION] *Coloro che effettuano scarichi esistenti di acque reflue, sono obbligati, fino al momento nel quale devono osservare i limiti di accettabilità stabiliti dal presente decreto, ad* ***adottare*** [*le misure necessarie ad evitare un aumento anche temporaneo dell'inquinamento* $_{Value}$]. [CNI $_{Agent}$]

b) [FINING] *Chiunque violi le norme tecniche e le modalità definite dal decreto di cui al comma 104, è soggetto alla* ***sanzione amministrativa*** [*del pagamento di una somma non inferiore a euro 100.000 e non superiore a euro 300.000* $_{Fine}$]. [CNI $_{Payer}$]

c) [COMPLIANCE] *Sono escluse dai procedimenti di deroga e sono comunque obbligate al* ***rispetto*** [*dei limiti previsti dalla normativa* $_{Norm}$] *le industrie alimentari ad eccezione di quelle di tipo artigianale con distribuzione del prodotto in ambito locale.* [CNI $_{Protagonist}$]

- situazioni descrittive di doveri 'fattuali generici', istanze di frames rappresentativi di realtà extragiuridiche generali, come esemplificato nei seguenti periodi:

d) [CAUSE_TO BE_INCLUDED] *Ove l'esame delle giustificazioni richieste e prodotte non sia sufficiente ad escludere l'incongruità della offerta, il concorrente è chiamato ad* ***integrare*** [*i documenti giustificativi* $_{New\_member}$] *ed all'esclusione potrà provvedersi solo all'esito della ulteriore verifica, in contraddittorio.* [CNI $_{Agent}$]

e) [COLLABORATION] *Gli enti gestori delle reti ed opere destinate al pubblico servizio in qualsiasi modo interferenti con l'infrastruttura da realizzare hanno l'obbligo di* ***cooperare*** [*alla realizzazione della stessa* $_{Undertaking}$] [*con le modalità previste dall'articolo 5* $_{Manner}$], *come precisato dal presente articolo.* [CNI $_{Partners}$]

f) [PARTICIPATION] *I produttori che non dimostrano di adottare adeguati provvedimenti sono obbligati a* ***partecipare*** [*ai consorzi di cui all'articolo 40* $_{Institution}$], *fatti salvi l'obbligo di corrispondere i contributi pregressi e l'applicazione delle sanzioni di cui all'articolo 54.* [CNI $_{Participants}$]

- situazioni descrittive di doveri 'fattuali specialistici', come esemplificato nei seguenti periodi:

g) [USING] *A decorrere dal 1 gennaio 2003, il tenore massimo di zolfo negli oli combustibili pesanti non può superare l'1.00 per cento in massa, fatti salvi i casi per i quali, ai sensi del decreto del Presidente del Consiglio dei Ministri 2 ottobre 1995, è obbligatorio l'****utilizzo*** [*di oli combustibili pesanti con un tenore massimo*

di zolfo non superiore allo 0.3 per cento in massa $_{Agent}$]. [CNI $_{Instrument}$]

h) [ACTIVITY_START] *Chi con il proprio comportamento omissivo o commissivo, in violazione delle disposizioni del presente decreto, provoca un danno alle acque, al suolo, al sottosuolo ed alle altre risorse ambientali, ovvero determina un pericolo concreto ed attuale di inquinamento ambientale, è tenuto a* ***procedere*** [*a proprie spese* $_{Manner}$] [*agli interventi di messa in sicurezza, di bonifica e di ripristino ambientale delle aree inquinate e degli impianti dai quali è derivato il danno, ovvero deriva il pericolo di inquinamento* $_{Activity}$]*, ai sensi e secondo il procedimento di cui all'articolo 17 del decreto legislativo 5 febbraio 1997, n. 22.* [CNI $_{Agent}$]

i) [DESTROYING] (*Gli accordi di programma di cui al comma 5 prevedono obbligatoriamente:*)[21] *c) lo* ***smaltimento*** [*delle sostanze lesive non rigenerabili né riutilizzabili* $_{Undergoer}$] [*, nel rispetto delle norme contro l'inquinamento e degli indirizzi emanati dal Ministro dell'ambiente con i regolamenti di cui al comma 7* $_{Manner}$]; [CNI $_{Destroyer}$]

A proposito di quest'ultimo tipo di 'doveri' è interessante far osservare che i frames presenti in FrameNet in grado di descrivere le situazioni 'fattuali specialistiche' regolate non sono specifici per il dominio ambientale oggetto di AMBnorm(Stato). Sono al contrario le istanze dei singoli FEs a renderli tali. Così, ad esempio, nel periodo g) la natura specialistica del dovere imposto è data dall'istanza del FE Instrument *di oli combustibili pesanti con un tenore massimo di zolfo non superiore allo 0.3 per cento in massa*, più che dal frame generico USING. Ciò è diretta conseguenza del fatto che FrameNet è stato pensato e sviluppato per la rappresentazione del significato di testi giornalistici rappresentativi della lingua comune.

In aggiunta all'osservazione appena fatta, è d'interesse qui mettere in evidenza che proprio per questo motivo i frames definiti in FrameNet non sempre consentono di rappresentare in maniera soddisfacente la realtà extragiuridica del dominio ambientale. È il caso, ad esempio, del seguente periodo, dove si fa riferimento al carattere obbligatorio di un'attività molto specialistica,

[21] In base alle scelte di segmentazione del testo in periodi descritte nel Paragrafo 3.3.1.1, questa parte del testo nella quale è istanziato il frame 'deontico' è contenuta nel periodo precedente, ma per chiarezza è stata qui riportata.

quella della *caratterizzazione di base* che consiste nella 'determinazione delle caratteristiche dei rifiuti, realizzata con la raccolta di tutte le informazioni necessarie per uno smaltimento finale in condizioni di sicurezza':

- [BEING_OBLIGATORY] [*La caratterizzazione di base* $_{Duty}$] *è* ***obbligatoria*** [*per ciascun tipo di rifiuti* $_{Responsible\_party}$] *ed è effettuata nel rispetto delle prescrizioni stabilite nell'allegato 1 al presente decreto.*

In questo caso, in FrameNet non è presente un frame che permetta di descrivere in maniera appropriata una tale situazione–tipo.

7.3.3.2 L'annotazione di 'permessi'

Le annotazioni delle istanze del FE State_of_affairs del frame PERMITTING hanno consentito di mettere in luce alcuni dei comportamenti 'permessi' in AMBnorm(Stato). Tra gli esempi più significativi vi sono:

- quelli relativi a situazioni 'giuridiche', come esemplificato nei seguenti periodi:

  a) *Allo scopo di diffondere la conoscenza ambientale e sensibilizzare l'opinione pubblica, in merito alle modifiche legislative conseguenti all'attuazione della presente legge, è autorizzata la* ***spesa*** *di 250.000 euro per l'anno 2004.*

  b) [HINDERING] *Al fine di permettere la prosecuzione degli investimenti nel settore dei trasporti di cui all'articolo 2, comma 5, della legge 18 giugno 199, n. 194, favorendo la riduzione delle emissioni inquinanti derivanti dalla circolazione di mezzi adibiti a servizi di trasporto pubblico locale, sono autorizzati* ***limiti*** [*di impegno* $_{Action}$] [*quindicennali* $_{Duration}$] [*pari a 30 milioni di euro per l'anno 2003 e a ulteriori 40 milioni di euro per l'anno 2004* $_{Degree}$]. [CNI $_{Hindrance}$]

  c) [DOCUMENTS] *L'*[***autorizzazione*** $_{Document}$][22] *è concessa per un periodo massimo di dieci anni a decorrere dalla data della prima iscrizione o del rinnovo dell'iscrizione del principio attivo negli elenchi predisposti in sede comunitaria secondo le procedure di cui agli articoli 27 e 28, della direttiva 98/8/CE e comunque per un*

[22] È da notare che in questi due casi la LU coincide con l'istanza di un FE.

periodo non superiore al termine fissato per il principio attivo nella predetta sede. [DNI $_{Right}$]

- quelli relativi a situazioni 'fattuali generiche', come esemplificato nei seguenti periodi:

  d) [PARTICIPATION] *È autorizzata la* ***partecipazione*** [*italiana* $_{Participants}$] [*al Fondo multilaterale per il Protocollo di Montreal per la protezione della fascia di ozono* $_{Institution}$].

  e) [CAUSE_CHANGE] *Tali aree devono ricadere all'interno del medesimo bacino idrografico nel quale è stata autorizzata la* ***trasformazione*** [*di coltura* $_{Entity}$]. [CNI $_{Agent}$]

- quelli relativi a situazioni 'fattuali specialistiche', come esemplificato nei seguenti periodi:

  f) *È permessa l'****immissione sul mercato*** *soltanto dei motori nuovi conformi ai requisiti della presente direttiva, siano essi già montati su macchine o no.*

  g) [CREATING] *Il permesso di ricerca e la concessione di* ***coltivazione***[*degli idrocarburi* $_{Created\_entity}$] [*in terraferma* $_{Place}$] *costituiscono titolo per la costruzione degli impianti e delle opere necessari, degli interventi di modifica, delle opere connesse e delle infrastrutture indispensabili all'esercizio, che sono dichiarati di pubblica utilità.* [INI $_{Creator}$]

  h) [CAUSE_FLUIDIC_MOTION] *f) concessioni di grandi* ***derivazioni*** [*di acqua* $_{Fluid}$] [*che interessino il territorio di più regioni e più bacini idrografici* $_{Area}$] *in assenza della determinazione del bilancio idrico* [INI $_{Agent}$]

Come fatto osservare nel paragrafo precedente, l'annotazione di situazioni specialistiche relative sia al dominio giuridico sia a quello della materia ambientale legislata ha messo in luce come il caratterere generale di FrameNet non sempre permetta di rappresentare il contenuto proposizionale dei periodi di AMBnorm(Stato).

Nel caso, ad esempio, del periodo a), la natura specialistica della situazione evocata dal sostantivo *spesa* relativa al 'complesso delle uscite di uno Stato' non è rappresentata da nessuno dei frames presenti in FrameNet. Il

frame COMMERCE_PAY[23] potrebbe essere quello che più si avvicina, ma prevede l'esistenza di uno scambio commerciale assente nella situazione–tipo evocata in a).

Così come è da notare che il carattere specialistico della realtà permessa nel periodo f) richiederebbe l'introduzione in FrameNet di un nuovo frame in grado di catturare la situazione evocata dall'unità lessicale predicativa polirematica *immissione sul mercato*, che si riferisce all'atto con il quale un prodotto viene reso disponibile sul mercato comunitario per la prima volta, cioè quando esso fuoriesce dalla fase di fabbricazione al fine di essere distribuito o utilizzato. Ad oggi, infatti, non esiste un frame in grado di descrivere una situazione specifica simile.

7.3.3.3 L'annotazione di 'divieti'

Tra i più significativi esempi di stati di cose o azioni proibite, annotati a partire dai frames PROHIBITING e DENY_PERMISSION, vi sono:

- quelle relative a situazioni legate al mondo del diritto, quali l'interdizione dalla capacità di esercitare le funzioni di amministratore, sindaco, direttore generale, ecc...:

  a) [LEADERSHIP] *e) non si trovino in stato di interdizione legale ovvero di interdizione temporanea dagli* [***uffici direttivi*** $_{Activity}$] *delle persone giuridiche e delle imprese;* [DNI $_{Leader}$] [INI $_{Governed}$]

- quelle relative a situazioni 'fattuali generiche', come esemplificato nei seguenti periodi:

  b) [ACTIVITY_START] *È fatto divieto ai soggetti di cui al comma 2, lettera a), della legge quadro, di* ***procedere*** [*ad estensioni dei lavori affidati in concessione al di fuori delle ipotesi consentite dalla direttiva 93/37/CEE* $_{Activity}$], *previo aggiornamento degli atti convenzionali sulla base di uno schema predisposto dal Ministro delle infrastrutture e dei trasporti.* [CNI $_{Agent}$]

[23] Il frame è così definito in FrameNet: "This frame involves Buyers paying Money for Goods. In this frame the Money is the direct object, and is mapped to the theme of the transfer."

c) [BUILDING] *È inoltre vietata per dieci anni, sui predetti soprassuoli, la* ***realizzazione*** [*di edifici nonché di strutture e infrastrutture finalizzate ad insediamenti civili ed attività produttive* $_{Created\_entity}$], *fatti salvi i casi in cui per detta realizzazione sia stata già rilasciata, in data precedente l'incendio e sulla base degli strumenti urbanistici vigenti a tale data, la relativa autorizzazione o concessione.* [INI $_{Agent}$]

d) [DISPERSAL] *È vietata la* ***diffusione*** [*dei dati e delle informazioni riservate di cui al comma 2* $_{Individuals}$] [*, da parte di chiunque ne venga a conoscenza per motivi attinenti al suo ufficio* $_{Agent}$].

- quelle relative a situazioni 'fattuali specialistiche', quali il commercio (periodo c)):

e) [COMMERCE_SCENARIO] *È fatto divieto di* ***commercializzare*** [*pile e accumulatori contenenti più dello 0,0005 per cento in peso di mercurio* $_{Goods}$], *anche nel caso in cui tali pile e accumulatori sono incorporati in apparecchi.*

f) [CAUSE_FLUIDIC_MOTION] *Fermo restando il divieto di* ***scarico*** *o di immissione diretta* [*di acque meteoriche* $_{Fluid}$] [*nelle acque sotterranee* $_{Goal}$], *ai fini della prevenzione di rischi idraulici ed ambientali, le acque meteoriche di dilavamento, le acque di prima pioggia e di lavaggio, le acque contaminate derivanti da spandimenti o da operazioni di estinzione di incendi delle aree esterne devono essere convogliate ed opportunamente trattate, ai sensi dell'articolo 39, comma 3, del decreto legislativo 11 maggio 1999, n. 152, e successive modificazioni;* [INI $_{Agent}$]

g) [CAUSE_CHANGE] *Ove non diversamente disposto dalle leggi regionali, è vietata la* ***conversione*** [*dei boschi governati o avviati a fustaia* $_{Entity}$] [*in boschi governati a ceduo* $_{Final\_category}$], *fatti salvi gli interventi autorizzati dalle Regioni ai fini della difesa fitosanitaria o di altri motivi di rilevante interesse pubblico.* [INI $_{Agent}$]

Come messo precedentemente in evidenza, la realtà extragiuridica contenuta in un periodo è resa esplicita non tanto dal frame che lo descrive quanto dalle istanze lessicali specialistiche dei suoi FEs. È il caso, ad esempio, della situazione contenuta nel periodo g), riconducibile alla materia ambientale

legislata sulla base dell'istanza specifica dell'entità modificata (FE Entity), *i boschi governati o avviati a fustaia*.

Infine, è interessante far notare che a differenza di quanto rilevato per i 'doveri' e i 'permessi', questo esperimento di annotazione ha messo in luce che raramente in AMBnorm(Stato) i 'divieti' sono relativi a situazioni che appartengono al mondo del diritto. Il corpus di annotazione si è invece rivelato particolarmente ricco di situazioni vietate riconducibili alla realtà extragiuridica ambientale.

7.4 La realizzazione linguistica dei FEs

Una delle principali finalità di questo caso di studio era quella di mettere in luce come l'annotazione semantica basata sui principi di FrameNet permetta di rendere esplicito il modo in cui la semantica del discorso giuridico è veicolata da costruzioni sintattiche specifiche della lingua del diritto. A questo scopo, la scelta di assumere come punto di partenza dell'annotazione semantica la struttura sintattica a dipendenze generata in modo automatico si è rivelata particolarmente vantaggiosa[24]. Essa ha consentito infatti di rendere esplicito come i FEs sono, non solo lessicalmente, ma anche sintatticamente realizzati in un periodo.

In quanto segue, sono pertanto riportati e discussi alcuni significativi esempi di annotazione con l'obiettivo di focalizzare l'attenzione su come alcuni dei più distintivi comportamenti sintattici del corpus AMBnorm(Stato), monitorati nel Capitolo 4, si riflettano nella rappresentazione del contenuto semantico dei suoi periodi.

7.4.1 La lunghezza delle relazioni di dipendenza

Come dimostrato nel Paragrafo 4.2.3, una delle caratteristiche sintattiche più evidenti dei testi giuridici è la grande lunghezza delle relazioni di dipendenza sintattica, calcolata sulla base del numero di tokens che intercorrono tra una testa sintattica e il suo dipendente. Il monitoraggio di una tale peculiarità ha infatti dimostrato che l'intero corpus di atti normativo–amministrativi (fatta accezione per la Costituzione italiana) esaminati contiene relazioni di dipendenza caratterizzate da una lunghezza media di 14,42 tokens, un valore pari quasi al doppio della lunghezza media delle relazioni in 2Par (7,71) e

[24]Vedi Paragrafo 7.2.

comunque molto superiore anche rispetto alla lunghezza riscontrata in Rep (8,80)[25].

Il tipo di annotazione semantica qui proposta permette di rendere esplicite le conseguenze che un tale comportamento sintattico ha nell'organizzazione del contenuto informativo. Ne è un esempio l'annotazione del seguente periodo:

a) [BEING_OBLIGATED] [*Chi con il proprio comportamento omissivo o commissivo, in violazione delle disposizioni del presente decreto, provoca un danno alle acque, al suolo, al sottosuolo ed alle altre risorse ambientali, ovvero determina un pericolo concreto ed attuale di inquinamento ambientale* $_{Responsible\_party}$], *è* ***tenuto*** [*a procedere a proprie spese agli interventi di messa in sicurezza, di bonifica e di ripristino ambientale delle aree inquinate e degli impianti dai quali è derivato il danno, ovvero deriva il pericolo di inquinamento* $_{Duty}$], [*ai sensi e secondo il procedimento di cui all'articolo 17 del decreto legislativo 5 febbraio 1997, n. 22* $_{Condition}$].

L'attenzione è posta sulla realizzazione sintattica della porzione di testo istanza del FE Responsible_party del frame BEING_OBLIGATED. Come (parzialmente) illustrato nella porzione di annotazione sintattica riportata nella Figura 7.4, la relazione 'subj' che lega la testa verbale *tenuto* e al suo soggetto, la cui testa sintattica è il pronome relativo *chi*, ha una lunghezza di ben 46 tokens. Il FE Responsible_party risulta così essere istanziato in un costrutto lungo 46 tokens.

In una prospettiva di interpretazione del testo, un tale comportamento sintattico rischia di ostacolare seriamente il processo di comprensione del periodo. Caratteristica rivelatrice di complessità sintattica, la grande distanza testa/dipendente in un periodo è infatti messa in stretta relazione con l'aumento dei costi cognitivi di comprensione[26]. In questo caso, la lunghezza della relazione di dipendenza 'subj' può rappresentare un ostacolo alla comprensione di 'chi è tenuto ad adempiere al dovere imposto'.

[25] Vedi Paragrafo 4.2.3.2.

[26] Vedi le riflessioni di Fiorentino (2007).

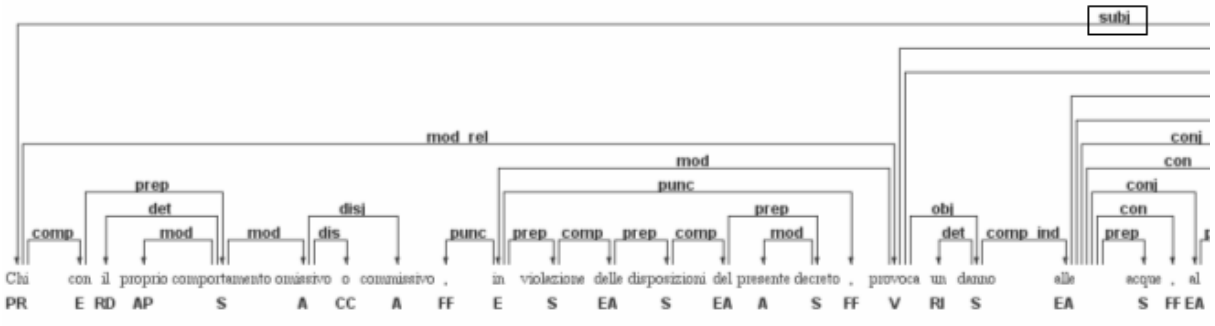


Figura 7.4: Un estratto della realizzazione sintattica dell'istanza del FE Responsible_party nel periodo a).

7.4.2 Il livello di incassamento gerarchico delle relazioni di dipendenza sintattica

Un'altro comportamento caratteristico dei testi giuridici esaminati riguarda la profondità degli incassamenti delle relazioni di dipendenza sintattica. La fase di monitoraggio ha infatti dimostrato che tali testi contengono incassamenti in media più profondi rispetto a quelli presenti nei testi giornalistici di riferimento[27].

A questa peculiarità sintattica corrisponde una particolare organizzazione dell'informazione. Un esempio significativo è rappresentato dal seguente periodo:

b) [BEING_OBLIGATED] [*In caso di mancato rispetto del programma di cui al comma 4, ovvero di mancata segnalazione ai sensi del comma 2* $_{Condition}$], [*il soggetto gestore* $_{Responsible\_party}$] [*ha* $_{Supp}$] *l'**obbligo*** [*di risarcire i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori* $_{Duty}$].

In questo caso, il FE Duty del frame BEING_OBLIGATED è istanziato nel testo dalla frase argomentale *di risarcire i danni subiti dal soggetto aggiudicatore per il conseguente impedimento al regolare svolgimento dei lavori*, dipendente dal token *obbligo* e costituita da una cascata di 8 relazioni di dipendenza consecutive, come mostra la Figura 7.5, dove l'area delle relazioni

[27] Vedi il Paragrafo 4.2.3.3.

incassate è delimitata da una linea tratteggiata e i singoli tipi di dipendenza ('arg', 'prep', 'obj', 'comp', ...) sono stati segnalati con una cornice rettangolare.

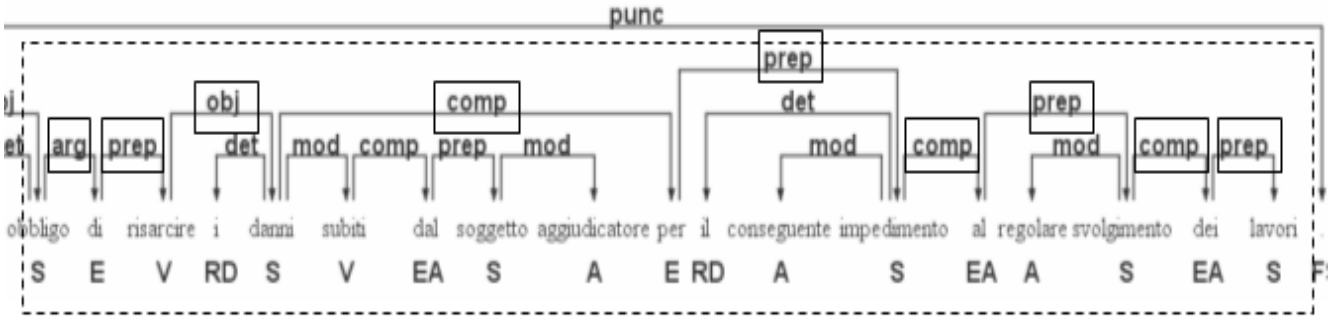


Figura 7.5: La realizzazione sintattica dell'istanza del FE Duty nel periodo b).

In stretta relazione con la lunghezza media dei periodi contenuti nei testi giuridici, maggiore rispetto a quella dei testi giornalistici di riferimento[28], un tale elevato numero di relazioni 'a cascata' è tra le cause responsabili della grande lunghezza (misurabile in tokens) delle istanze dei FEs. Sebbene infatti sia attualmente in corso uno studio della differenza tra la realizzazione sintattica dei FEs in testi giornalistici e giuridici, ad una prima analisi a campione risulta che questi ultimi contengono istanze mediamente più corte.

7.4.3 Le 'catene' di complementi preposizionali

In fase di monitoraggio linguistico, la spiccata propensione per la modificazione nominale era risultata essere una delle caratteristiche più significative degli atti normativo–amministrativi analizzati. Rispetto a questo parametro di monitoraggio, i testi giuridici avevano in particolare dimostrato di contenere incassamenti gerarchici di complementi preposizionali, modificatori di sostantivi, molto profondi, con sequenze 'a cascata' in media più lunghe dei testi giornalistici di riferimento[29].

[28] Vedi Paragrafo 4.2.1.
[29] Vedi Paragrafo 4.2.3.5.

La fase di annotazione semantica ha permesso di offrire una dimostrazione di come un tale comportamento sintattico sia da annoverarsi tra i maggiori responsabili delle ben note "complicazioni strutturali" proprie dei testi giuridici, come fatto osservare da Garavelli (2001, p. 175).

È il caso dell'esempio che segue, dove l'informazione relativa alla data (FE Time) a partire dalla quale il soggetto è obbligato a soddisfare l'obbligo è annidata in una 'catena' di 10 incassamenti preposizionali a cominciare dal token *data*, come mostra l'estratto di annotazione sintattica riportato nella Figura 7.6:

c) [BEING_OBLIGATED] [*A decorrere dalla data di scadenza del termine di novanta giorni dalla data di pubblicazione nella Gazzetta Ufficiale del decreto di approvazione dello Statuto di cui al comma 2* $_{Time}$], [*chiunque, in ragione della propria attività, detiene oli e grassi vegetali e animali esausti* $_{Responsible\_party}$] *è* ***obbligato*** [*a conferirli al Consorzio direttamente o mediante consegna a soggetti incaricati del Consorzio* $_{Duty}$].

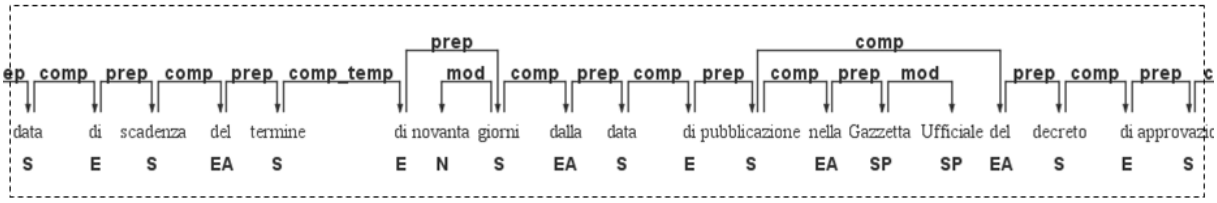


Figura 7.6: Un estratto della realizzazione sintattica dell'istanza del FE 'Time' nel periodo c).

7.4.4 Le dipendenze di predicati verbali

Nel Paragrafo 4.2.3.4, era stata posta l'attenzione su di un'altra caratteristica sintattica dei testi giuridici: la presenza di un numero medio di dipendenti da teste verbali, fossero essi di tipo nominale, argomenti sottocategorizzati o modificatori di varia natura (locativi, temporali, causali, ecc...), inferiore ai testi giornalistici di riferimento.

Tra gli altri, uno dei corpora con i valori più bassi era risultato essere proprio AMBnorm(Stato), con una percentuale di teste verbali con un numero

medio di dipendenti uguale a 2, pari al 22,20% del totale di teste verbali considerate[30].

In quell'occasione era stata avanzata l'ipotesi che le ragioni di questa differenza fossero riconducibili alla maggiore presenza di forme participiali e di forme verbali passive rispetto ai testi giornalistici di riferimento[31]. In fase di annotazione semantica è stato verificato come questa intuizione potesse avere un riscontro effettivo nell'organizzazione del contenuto informativo di un periodo.

Un esempio è rappresentato dall'annotazione del seguente periodo:

d) [PERMITTING] *Al fine di permettere la prosecuzione degli investimenti nel settore dei trasporti di cui all'articolo 2, comma 5, della legge 18 giugno 1998, n. 194, favorendo la riduzione delle emissioni inquinanti derivanti dalla circolazione di mezzi adibiti a servizi di trasporto pubblico locale, sono* ***autorizzati*** [*limiti di impegno quindicennali pari a 30 milioni di euro per l'anno 2003 e a ulteriori 40 milioni di euro per l'anno 2004* $_{State\_of\_affairs}$]. [CNI $_{Principle}$]

In questo caso, la presenza della forma passiva riduce da tre a due i dipendenti del verbo *autorizzare*. Come mostra, infatti, la Figura 7.7, il token *autorizzato* ha due dipendenti (esclusa la punteggiatura): *i)* la subordinata implicita *Al fine di permettere la prosecuzione degli investimenti nel settore dei trasporti di cui all'articolo 2, comma 5, della legge 18 giugno 1998, n. 194, favorendo la riduzione delle emissioni inquinanti derivanti dalla circolazione di mezzi adibiti a servizi di trasporto pubblico locale*, legata da una relazione di tipo 'arg'[32] e *ii)* il soggetto passivo, il sotto–albero cioè la cui testa sintattica *limiti* è legata da una relazione di tipo 'subj_pass' alla testa verbale *autorizzato* (radice sintattica dell'intero periodo), come segnalato dalla cornice rettangolare.

[30]La percentuale di teste verbali con un numero medio di dipendenti uguale a 2 in Rep è del 32,33% e in 2Par è del 33,50%.

[31]Vedi Paragrafo 4.2.2.1.

[32]Per chiarezza si ricorda qui che, sulla base dello schema di annotazione sintattica a dipendenze, la relazione 'complement' ('comp') è la relazione tra una testa e un complemento preposizionale, sia esso modificatore o argomento. Questa relazione funzionale sottospecificata è particolarmente utile in quei casi in cui è difficile stabilire la natura argomentale o di modificatore del complemento. Per ragioni di spazio la dipendenza non è mostrata nella Figura 7.7.

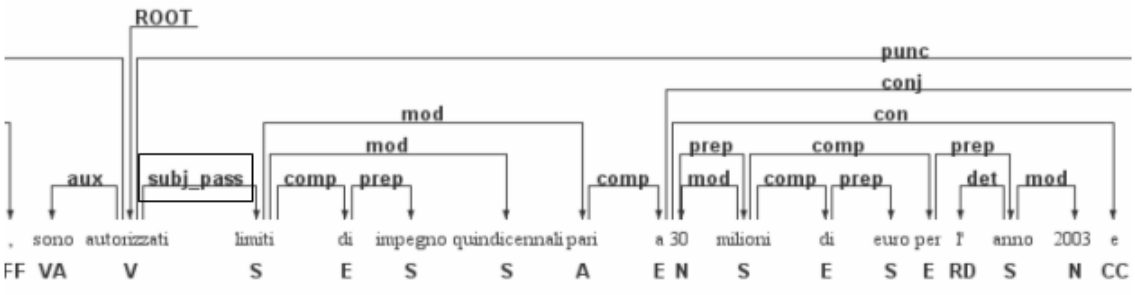


Figura 7.7: Uno dei due dipendenti della testa verbale *autorizzato* nel periodo d).

La realizzazione sintattica del verbo non consente dunque che sia linguisticamente espresso 'chi' autorizza lo stato di cose permesso. Tuttavia, l'annotazione di un caso di 'Constructional Null Instantiation' (CNI) permette di rendere esplicito che per la piena comprensione della situazione–tipo evocata dal verbo *autorizzare* è necessario presupporre la presenza di un FE Principle, non lessicalmente e sintatticamente espresso.

Un secondo esempio di come il modello di annotazione semantica basato sui principi di FrameNet consenta di mettere in luce in che modo il numero ridotto di dipendenti da teste verbali abbia conseguenze sull'organizzazione del contenuto semantico–lessicale dei testi giuridici è rappresentato dal seguente periodo:

e) [IMPOSING_OBLIGATION] *È* [*fatto* $_{Supp}$] ***obbligo*** [*a chiunque spetti* $_{Responsible\_party}$] [*di osservarlo e di farlo osservare* $_{Duty}$]. [CNI $_{Obligator}$]

In questo caso, il frame IMPOSING_OBLIGATION è evocato dalla costruzione a verbo supporto *fare obbligo*. Oltre che dalla forma verbale passiva, l'annotazione semantica è condizionata anche dalla scelta lessicale. Se fosse stato scelto il verbo *obbligare*, la sua struttura valenziale avrebbe implicato la presenza di tre dipendenti (soggetto, oggetto, complemento). Al contrario, la scelta del sostantivo *obbligo* e del costrutto participiale passivo *è fatto* ha reso implicato l'omissione della realizzazione lessicale dell''agente che obbliga' (FE Obligator). La sua presenza è per questo motivo stata segnalata come un caso di 'Costructional Null Instantiation' (CNI), di omissione dovuta appunto alla struttura stessa del periodo.

7.5 I diversi aspetti dell'Obligation_scenario

Come fatto notare nel Paragrafo 6.5.2, una delle principali potenzialità di FrameNet come modello di riferimento per la rappresentazione ontologica della conoscenza di dominio riguarda la possibilità di descrivere una determinata situazione–tipo assumendo prospettive diverse. Questo consente di rappresentare un medesimo scenario conoscitivo adottanto i diversi punti di vista delle varie entità coinvolte.

È il caso, ad esempio, dello scenario relativo all'**obbligo**, che nelle 'core legal ontologies' è rappresentato come un concetto unitario, corrispondente ad una classe ontologica messa in relazione con le altre classi dell'ontologia grazie ad una serie di relazioni. Sono qui riportati due casi esemplificativi di questo stato di cose.

Come si può vedere nella Figura 7.8, nella "Core Legal Ontology" (CLO)[33], ad esempio, la classe 'Obligation'[34] è legata da una relazione di tipo 'is–a' (subClassOf) alla classe 'LegalModalDescription'[35], a sua volta legata dallo stesso tipo di relazione gerarchica alla classe 'LegalDescription'[36].

Seppure inserito in una rete di relazioni ontologiche modellata su principi teorici diversi, anche nel caso di una seconda ontologia giuridica qui considerata, la "LKIF–Core ontology" (Breuker et al., 2007), il concetto di **obbligo** è rappresentato come un'entità atomica. Come mostrato nella Figura 7.9, la classe 'Obligation' è legata da una relazione di tipo 'is–a' (equivalentClassOf) alla classe 'Prohibition'[37], ulteriormente specificata dalla relazione gerarchica 'subClassOf' che la lega alla classe 'Norm'.

Al contrario, in FrameNet l'obbligo è descritto *i)* dal 'Non–lexical frame' Obligation_scenario, un frame cioè per il quale non sono state previste LUs evocatrici né annotazioni e che ha unicamente la funzione di mettere in

[33]http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl

[34]La classe è descritta dalla glossa "the proposition expressing the obligation to perform a certain action is true whenever optimal practical cognition would lead one to have the intention of accomplishing that action".

[35]La classe è definita come "The set of normative positions from Hohfeld's works (and his continuators)".

[36]La classe è definita come "A social description having legal validity and possibly effects. They can be either legal norms, principles, rationales, contracts, regulations to enforce norms, etc.".

[37]La classe è qui accompagnata infatti dalla glossa "Prohibition obliges/allows thing(s), to which therefore the predication Obliged applies, and disallows thing(s), to which therefore the predication Disallowed applies".

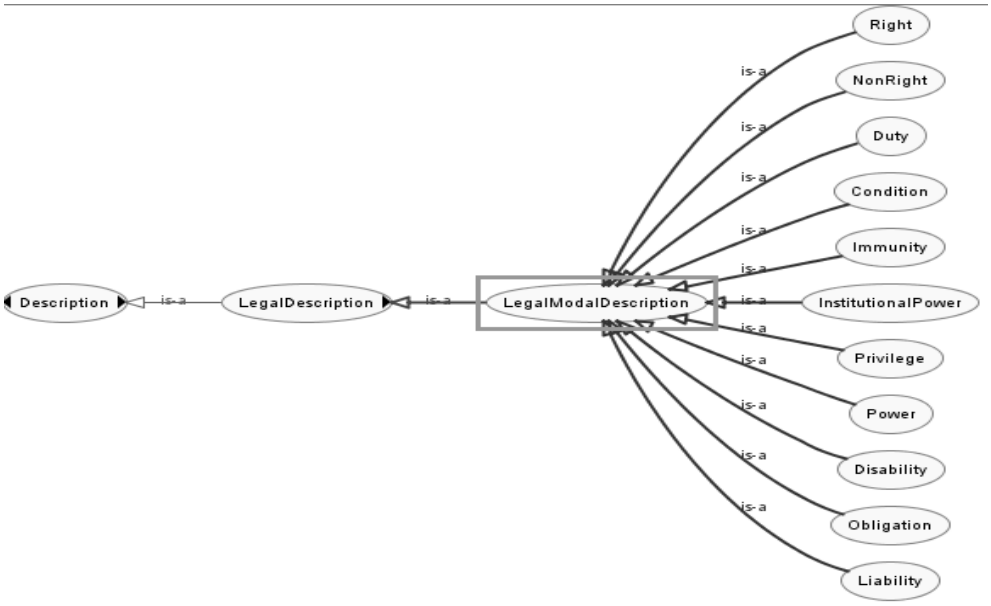


Figura 7.8: La rappresentazione del concetto di **obbligo** nella "Core Legal Ontology".

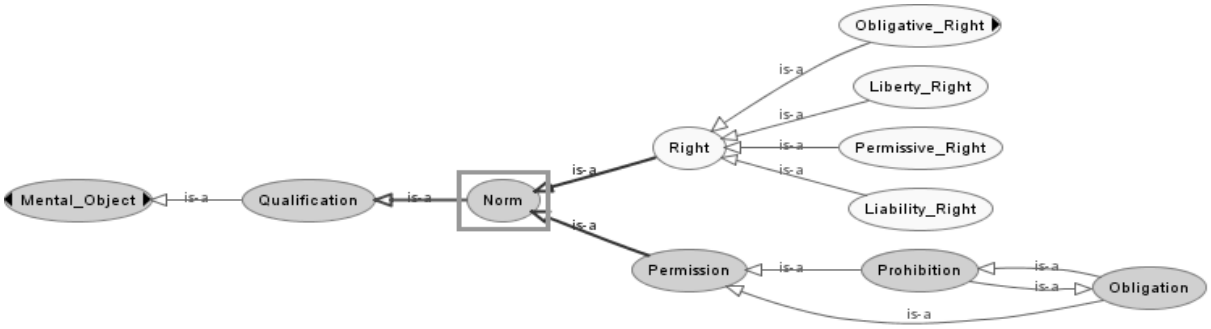


Figura 7.9: La rappresentazione del concetto di **obbligo** nella LKIF–Core ontology.

collegamento due o più frames (Ruppenhofer et al., 2010, p. 80), e *ii)* dalla rete di relazioni 'frame–to–frame' che legano questo 'Super frame' ai diversi 'Sub frames' espressione dei seguenti punti di vista prospettici dai quali è

possibile guardare questo scenario[38]:

a) il punto di vista della 'parte obbligata' (descritto nel frame BEING_OBLIGATED);

b) il punto di vista del 'dovere' da adempiere (descritto nel frame BEING_OBLIGATORY);

c) il punto di vista di chi prescrive l'obbligo (descritto nel frame IMPOSING_OBLIGATION);

d) l'aspetto di conformità di un'azione rispetto al 'dovere' imposto (descritto nel frame COMPLIANCE);

e) lo stato operativo (o non operativo), di vigenza, cioè di mera esistenza della norma che obbliga all'interno dell'ordinamento giuridico (descritto nel frame BEING_IN_EFFECT);

f) lo status giuridico di un documento, di cui l'obbligare è uno degli stati possibili (descritto nel frame DOCUMENTS);

g) nonché, la dimensione 'anankastica' di un comportamento richiesto, rispetto alla quale lo svolgere un'azione è una 'condizione necessaria' ma non deonticamente prescrittiva che regola uno stato di cose (descritta nel frame REQUIRED_EVENT).

In quanto segue è discusso dunque quali siano le singole relazioni 'frame–to–frame' a permettere di assumere una visione prospettica sul frame OBLIGATION_SCENARIO.

7.5.1 La relazione Perspective_on

La relazione Perspective_on[39] che lega i due 'Sub frames' BEING_OBLIGATED e BEING_OBLIGATORY al 'Super frame' OBLIGATION_SCENARIO consente di avere due punti di vista diversi e complementari sul medesimo scenario. I due 'Perspectivized frames' permettono dunque di descriverlo da due prospettive: quella di chi è obbligato ad adempiere l'obbligo e quella del 'dovere' imposto.

[38] Per la rappresentazione grafica della rete di relazioni 'frame–to–frame' vedi la Figura 7.1.

[39] La relazione è definita da Ruppenhofer et al. (2010, p. 75) come "a refinement of the more general Using relation. Perspective on constrains related frames considerably more. The use of this relation indicates the presence of at least two different points–of–view that can be taken on the Neutral frame".

Tale comune visione prospettica è organizzata a livello dei singoli FEs come illustrato nella Tabella 7.10[40]. I due 'Core' FEs Duty e Responsible_party corrispondono, mentre c'è qualche variazione per quanto riguarda i 'Non–Core' FEs, con le conseguenze discusse nel Paragrafo 7.6.1.

BEING_OBLIGATORY	OBLIGATION_SCENARIO	BEING_OBLIGATED
Condition	Condition	Condition
Consequence	Consequence	Consequence
	Degree	
Duty	**Duty**	**Duty**
		Frequency
Explanation		
Place	Place	Place
Purpose	Purpose	
Responsible_party	**Responsible_party**	**Responsible_party**
Time	Time	Time

Tabella 7.10: Confronto tra i FEs dei frames BEING_OBLIGATED, OBLIGATION_SCENARIO e BEING_OBLIGATORY.

7.5.2 La relazione Causative_of

L'aspetto causativo della prospettiva assunta nel frame IMPOSING_OBLIGATION è espresso dalla relazione di tipo Causative_of che lega questo frame al frame BEING_OBLIGATED. Si tratta di una "fairly systematic non–inheritance relationship between stative frames and the causative frames which refer to them" (Ruppenhofer et al., 2010, p. 77).

Come illustrato nel seguente esempio di annotazione, la situazione descritta dal frame causativo prevede la presenza di un soggetto giuridico che impone l'obbligo su un soggetto tenuto ad adempierlo:

- [IMPOSING_OBLIGATION] [*Salvo che il fatto costituisca reato* $_{Condition}$], [*le sanzioni di cui ai commi 1 e 2* $_{Duty}$] *sono* ***irrogate*** [*dalle autorità competenti* $_{Obligator}$], [*sulla base degli accertamenti effettuati dalle autorità abilitate ai controlli ai sensi dell'articolo 14* $_{Situation}$].

A livello delle corrispondenze tra i singoli FEs dei due frames coinvolti nella relazione, come si può vedere nella Tabella 7.11, il ruolo causativo è infatti

[40] In grassetto sono segnati i 'Core' FEs.

svolto da un soggetto obbligante (FE Obligator) o da un principio regolativo (FE Principle), due FEs specifici del frame IMPOSING_OBLIGATION.

IMPOSING_OBLIGATION	BEING_OBLIGATED
Condition	Condition
	Consequence
Duty	**Duty**
Manner	
Means	
Obligator	
Place	Place
	Frequency
Principle	
Purpose	
Responsible_party	**Responsible_party**
Situation	
Time	Time

Tabella 7.11: Confronto tra i FEs dei frames BEING_OBLIGATED e IMPOSING_OBLIGATION.

7.5.3 La relazione Using

I frames legati da questo tipo di relazione al 'Non–lexical frame' OBLIGATION_SCENARIO sono tre: COMPLIANCE, BEING_IN_EFFECT e DOCUMENTS. Si tratta di una relazione molto generale in base alla quale "a particular frame makes reference in a very general kind of way to the structure of a more abstract, schematic frame" (Ruppenhofer et al., 2010, p. 78).

Come si può vedere nella Tabella 7.12, che riporta la corrispondenza tra i frames a livello dei singoli FEs, la relazione Using permette di rendere esplicito che il 'dovere' che un soggetto è obbligato ad adempiere (FE Duty in OBLIGATION_SCENARIO) è l'azione che deve essere conforme alla legge (FE Act in COMPLIANCE), nonché lo status del contenuto di un documento con valore giuridico (FE Obligation in DOCUMENTS) che obbliga il suo possessore a svolgere una certa azione o a tenere un certo comportamento. Di conseguenza, il soggetto che deve adempiere l'obbligo (FE Responsible_party in OBLIGATION_SCENARIO) è il soggetto il cui comportamento deve essere conforme alla legge (FE Protagonist in COMPLIANCE), nonché il soggetto obbligato a seguire le regole di comportamento vigenti (FE Obligated_party

Obligation_scenario	Compliance	Being_in_effect	Documents
Condition			Specification
Consequence			
Place		Place	
Purpose			
Duty	**Act**		**Obligation**
			Descriptor
Degree	Degree	Degree	
			Issuer
			Medium
			Right
			Status
			Document
	Depictive		
	Judge		
	Manner		
	Means		
		Binding_principle	
		Circumstances	
		Duration	
		Explanation	
	Norm		
Responsible_party	**Protagonist**	**Obligated_party**	**Bearer**
	Reason		
	Result		
	State of affairs		
Time	Time		

Tabella 7.12: Confronto tra i FEs dei frames Obligation_scenario, Compliance, Being_in_effect e Documents.

in Being_in_effect) e il possessore del documento che certifica la realtà dell'obbligo (FE Bearer in Documents).

7.5.4 La relazione Inheritance

La relazione che lega i frames Obligation_scenario e Required_event è il tipo di relazione più forte in FrameNet, equivalente alla relazione ontologica di tipo 'is–a'. Questo comporta che "anything which is strictly true about the semantics of the Parent must correspond to an equally or more specific fact about the Child" (Ruppenhofer et al., 2010, p. 75).

OBLIGATION_SCENARIO	REQUIRED_EVENT
Condition	Circumstances
Consequence	Negative_consequences
Place	Place
Purpose	Purpose
Duty	**Required_situation**
Degree	Degree
Responsible_party	
	Explanation
Time	Time

Tabella 7.13: Confronto tra i FEs dei frames OBLIGATION_SCENARIO e REQUIRED_EVENT.

In base all'organizzazione della rete di frames prevista in FrameNet, è il frame OBLIGATION_SCENARIO il 'Child frame' che eredita i FEs del 'Parent frame' REQUIRED_EVENT. Come si può vedere nella Tabella 7.13, che riporta la relazione tra i due frames a livello dei singoli FEs, il 'Child frame' eredita quasi tutti i FEs del 'Parent frame'.

L'eccezione più significativa è rappresentata dal FE Responsible_party, presente soltanto nel frame OBLIGATION_SCENARIO. Degno di nota è inoltre il fatto che il FE Duty del frame OBLIGATION_SCENARIO corrisponda al FE Required_situation nel frame REQUIRED_EVENT.

Entrambe le differenze sono riconducibili al diverso valore illocutivo degli enunciati espressione delle due situazioni–tipo considerate. Rifacendosi alla distinzione operata in ambito linguistico da Conte (1995), si può affermare che il frame OBLIGATION_SCENARIO rappresenti la modalità propriamente **deontica** di un enunciato, mentre il frame REQUIRED_EVENT ne rende esplicito il valore **anankastico**.

Ciò è evidente proprio a livello delle differenze tra i FEs presenti nei due frames. Coerentemente con il fatto che una prescrizione deontica deve sempre avere un destinatario, nel frame 'deontico' è prevista la presenza di un Responsible_party al quale è indirizzato l'obbligo; mentre nel frame 'anankastico' un tale elemento manca.

La differenza fondamentale che Maria–Elisabeth Conte (1995) rileva nel descrivere la natura degli enunciati anankastici riguarda il fatto che essi "non qualificano deonticamente (come obbligatorio, come vietato, come permesso, come indifferente) un comportamento, ma pongono una condizione ne-

cessaria (positiva o negativa) di qualcosa (d'un atto, d'uno stato di cose, d'un oggetto)". È interessante qui osservare come ciò sia catturato dalle situazioni–tipo descritte dai due frames in esame: il comportamento o l'azione in esame svolge il ruolo di 'dovere che deve essere obbligatoriamente svolto' (FE Duty) nel frame OBLIGATION_SCENARIO e di 'stato di cose necessario' (FE Required_situation) nel frame REQUIRED_EVENT.

Dal momento che le due modalità hanno realizzazione linguistiche simili, la soluzione proposta da Conte (1995) per distinguere i due casi è quella del "test della ripresa anaforica". Se cioè l'enunciato ammette una prosecuzione tramite la ripresa dell'azione imposta/richiesta "quest'obbligo" allora si tratta di un enunciato con valore deontico; al contrario se l'azione può essere anaforicamente ripresa utilizzando l'espressione "questo requisito" allora l'enunciato ha valore anankastico. Ciò è reso possibile grazie al fatto che, come fa notare Maria–Elisabeth Conte stessa, "le due prosecuzioni rimandano a due differenti contesti d'uso".

Durante gli esperimenti di annotazione condotti in questo studio, è stato sperimentato se e come fosse possibile rintracciare in AMBnorm(Stato) le due modalità deontica e anankastica. Un esempio significato è quello rappresentato dalla coppia di periodi che seguono[41]:

a) [IMPOSING_OBLIGATION] [*In conformità alla vigente normativa in materia di smaltimento dei rifiuti* $_{Condition}$], [*è fatto* $_{Supp}$] ***obbligo*** [*a tutti i detentori di prodotti, di impianti e di beni durevoli contenenti le sostanze lesive* $_{Responsible\_party}$] [*di conferire i medesimi, al termine della loro durata operativa, a centri di raccolta autorizzati* $_{Duty}$]. [CNI $_{Obligator}$]

b) [REQUIRED_EVENT] [*Per il rilascio dell'autorizzazione* $_{Purpose}$], [*ai fini della verifica della conformità urbanistica dell'opera* $_{Explanation}$], [*è fatto* $_{Supp}$] ***obbligo*** [*di richiedere il parere motivato degli enti locali nel cui territorio ricadano le opere di cui al comma 1* $_{Required\_event}$].

Sebbene, dunque, i due periodi condividano la medesima LU *obbligo*, essa tuttavia è evocatrice di un diverso contenuto informativo.

[41] È da notare che, dal momento che il frame OBLIGATION_SCENARIO è un 'Non–lexical frame', è stato necessario considerare il caso di un frame deontico ad esso legato.

7.6 Proposte di specializzazioni di dominio

Le proposte di specializzazione elaborate in questo studio sulla scorta degli esperimenti di annotazione condotti sono state classificate rispetto ai diversi livelli di descrizione semantica coinvolti.

7.6.1 Specializzazioni di FEs

In questo caso, i due tipi di modifiche di FrameNet individuati riguardano:

- la specializzazione di FEs già contenuti in un frame, allo scopo di descrivere in modo più preciso lo specifico ruolo svolto da entità parte di una determinata situazione–tipo;
- l'aggiunta ex novo di FEs non ancora presenti in FrameNet, con lo scopo di rendere più esaustiva la descrizione di tutti gli elementi conoscitivi necessari alla completa rappresentazione della conoscenza di dominio.

Sono di seguito discussi alcuni esempi.

7.6.1.1 Specializzazione di FEs già esistenti

Un caso significativo è quello costituito dalla specializzazione dell'informazione relativa alla descrizione delle 'circostanze' normative nelle quali si svolge uno scenario deontico.

Come esempio di questo tipo di specializzazione di dominio è stato scelto il caso delle circostanze di realizzazione di uno scenario di 'obbligo'. Si è ritenuto che fosse importante specificare il ruolo particolare svolto dal generico FE Condition presente nei frames Being_obligated, Being_obligatory e Imposing_obligation.

Sebbene al momento tutte le istanze sono state annotate come realizzazioni di un unico FE Condition, è stata tuttavia condotta una rassegna dei diversi tipi di 'circostanze di realizzazione dell'obbligo' finalizzata a raccogliere esempi sufficienti per una futura specializzazione di dominio di FrameNet in questo senso.

Come si può vedere negli esempi che seguono, la questione è strettamente collegata con il dibattito relativo alla codifica della semantica dei connettivi linguistici nei testi giuridici. Processo fondamentale per la completa rappresentazione della struttura logico–concettuale del testo, è il tema centrale

di discussione di Visconti (2009). È questo motivo per cui la classificazione dei diversi tipi di 'condizioni' rintracciati negli esperimenti di annotazione è stata condotta sulla base delle diverse tipologie di connettivi individuate da Visconti (2009).

In quanto segue, le istanze del FE Condition annotate sono dunque state classificate in base al loro diverso contributo alla semantica del periodo[42]:

- 'condizioni' che hanno la funzione di delimitare il campo dell'obbligo facendo per lo più riferimento alla normativa che lo regola, es.

  a) [BEING_OBLIGATED] [*Le attività di trasporto e dispacciamento del gas naturale a rete, nonché la gestione di infrastrutture di approvvigionamento di energia connesse alle attività di trasporto e dispacciamento di energia a rete* $_{Responsible\_party}$], *sono di interesse pubblico e* [*sono sottoposte* $_{Supp}$] *agli* ***obblighi*** [*di servizio pubblico* $_{Duty}$] [*derivanti dalla normativa comunitaria, dalla legislazione vigente e da apposite convenzioni con le autorità competenti* $_{Condition}$].

  b) [IMPOSING_OBLIGATION] [*In conformità alla vigente normativa in materia di smaltimento dei rifiuti* $_{Condition}$], [*è fatto* $_{Supp}$] ***obbligo*** [*a tutti i detentori di prodotti, di impianti e di beni durevoli contenenti le sostanze lesive* $_{Responsible\_party}$] [*di conferire i medesimi, al termine della loro durata operativa, a centri di raccolta autorizzati* $_{Duty}$]. [CNI $_{Obligator}$]

  c) [BEING_OBLIGATORY] [*Il procedimento di valutazione di impatto ambientale* $_{Duty}$] *è* ***obbligatorio*** *e vincolante* [*per tutte le opere ad esso soggette* $_{Responsible\_party}$] [*a norma delle vigenti disposizioni* $_{Condition}$].

- 'condizioni' che hanno valore di 'preservato' che dunque subordinano l'adempimento dell'obbligo ad uno stato di cose preesistente e sempre valido, es.

  d) [BEING_OBLIGATED] [*I produttori che non dimostrano di adottare adeguati provvedimenti* $_{Responsible\_party}$] *sono* ***obbligati*** [*a partecipare ai consorzi di cui all'articolo 40* $_{Duty}$], [*fatti salvi l'obbligo di*

[42]Per chiarezza è riportata l'intera annotazione del periodo.

corrispondere i contributi pregressi e l'applicazione delle sanzioni di cui all'articolo 54 $_{Condition}$].

e) [BEING_OBLIGATED] [*Il contraente generale* $_{Responsible\_party}$] [*assume* $_{Supp}$] *l'**obbligo*** [*di verificare il progetto esecutivo posto in gara e di farlo proprio* $_{Duty}$], [*fermo restando quanto disposto dal comma 5 dell'articolo 9* $_{Condition}$].

- 'condizioni' che fanno riferimento ad un disposto che è in contrasto con l'obbligo da adempiere, es.

f) [BEING_OBLIGATED] [*In deroga all'articolo 30, comma 2, della legge 11 febbraio 1994, n. 109, e successive modificazioni* $_{Condition}$], [*l'esecutore dei lavori* $_{Responsible\_party}$] *è **obbligato*** [*a costituire una garanzia fidejussoria, da parte di un istituto di credito di primaria importanza a livello nazionale, del 50 per cento dell'importo degli stessi, destinata a garantire l'ultimazione dell'opera entro il termine fissato dal bando di gara* $_{Duty}$].

- 'condizioni' con valore esclusivo, es.

g) [BEING_OBLIGATED] [*I titolari degli impianti di incenerimento* $_{Responsible\_party}$] *sono **obbligati*** [*ad accettare il predetto materiale e le predette proteine animali* $_{Duty}$] [*salvo che, nell'ipotesi di materiale specifico a rischio tal quale, siano esonerati dalle regioni o province autonome competenti per riconosciuta inidoneità degli impianti stessi* $_{Condition}$].

- 'condizioni' con valore condizionale, es.

h) [BEING_OBLIGATED] [*Qualora, in attuazione delle disposizioni del comma 2, siano avviate al consumo in rete miscele combustibile diesel-biodiesel con contenuto in biodiesel in misura superiore al 5 per cento* $_{Condition}$], [*i punti vendita nei quali tali miscele sono distribuite* $_{Responsible\_party}$] *sono **obbligati*** [*ad esporre idonee etichette di descrizione del prodotto, unitamente all'elenco dei veicoli omologati per l'uso dei predetti biocarburanti* $_{Duty}$].

i) [BEING_OBLIGATED] [*Ove l'esame delle giustificazioni richieste e prodotte non sia sufficiente ad escludere l'incongruità della offerta* $_{Condition}$], [*il concorrente* $_{Responsible\_party}$] *è **chiamato*** [*ad integrare*

i documenti giustificativi $_{Duty}$] *ed all'esclusione potrà provvedersi solo all'esito della ulteriore verifica, in contraddittorio.*

È importante qui far notare come i diversi tipi di 'condizioni' abbiano anche realizzazioni linguistiche diverse. In un'ottica di estensione della collezione di periodi annotati semanticamente, questo apre la strada ad una futura identificazione automatica di FE(s) Condition specifici per il dominio.

7.6.1.2 Aggiunte ex novo di FEs

Un esempio significativo di questo tipo di specializzazione riguarda l'aggiunta di un FE Purpose ai frames BEING_OBLIGATED e PERMITTING. Per entrambi in FrameNet non è infatti prevista l'esistenza di questo FE. Tuttavia dall'analisi delle frasi annotate risulta chiaro come la 'finalità per cui un soggetto obbligato adempie un dovere' o 'per cui uno stato di cose viene permesso' sia un elemento informativo rilevante per la completa descrizione della situazione–tipo descritta.

Come esemplificato nei seguenti periodi, un FE Purpose sarebbe dunque necessario per rendere esplicito il proposito per cui *i concessionari* sono sottoposti all'obbligo di appalto e la finalità per la quale sono autorizzati *limiti di impegno*:

a) [BEING_OBLIGATED] [*Per la realizzazione delle opere previste nelle convenzioni già assentite alla data del 30 giugno 2002, ovvero rinnovate e prorogate ai sensi della legislazione vigente* $_{PURPOSE}$], [*i concessionari* $_{Responsible\_party}$] *sono* ***tenuti*** [*ad appaltare a terzi una percentuale minima del 40 per cento dei lavori* $_{Duty}$], [*applicando le disposizioni della presente legge ad esclusione degli articoli 7, 14, 19, commi 2 e 2-bis, 27, 32, 33* $_{Condition}$].

b) [PERMITTING] [*Al fine di permettere la prosecuzione degli investimenti nel settore dei trasporti di cui all'articolo 2, comma 5, della legge 18 giugno 1998, n. 194, favorendo la riduzione delle emissioni inquinanti derivanti dalla circolazione di mezzi adibiti a servizi di trasporto pubblico locale* $_{PURPOSE}$], *sono* ***autorizzati*** [*limiti di impegno quindicennali pari a 30 milioni di euro per l'anno 2003 e a ulteriori 40 milioni di euro per l'anno 2004* $_{State\_of\_affairs}$]. [CNI $_{Principle}$]

7.6.2 Specializzazioni di Semantic Types

La proposta di specializzazione rispetto a questo livello di rappresentazione del significato è in linea con l'approccio seguito dagli utilizzatori di FrameNet particolarmente attenti a metterne in evidenza l'aspetto di 'rete organizzata della conoscenza'. Come discusso nel Paragrafo 6.1.3, è attivo, infatti, un filone di ricerche finalizzato a collegare i singoli FEs con classi di alcune delle principali ontologie formali oggi esistenti.

Sulla scia di questi studi, le proposte qui avanzate sono motivate *i)* dalla ben nota consapevolezza che in FrameNet la gerarchia di STs è piuttosto ridotta; non tutti i FEs sono infatti arricchiti con informazione relativa alla restrizione di selezione semantica delle loro istanze lessicali; *ii)* dall'intuizione espressa da Scheffczyk et al. (2006a) di "constrain the filler types of FEs for specific domains" allo scopo di "help semantic parsers both with word sense disambiguation of predicators and identifying which pieces of a sentence fill FEs".

A questo si aggiunge il fatto che le ontologie giuridiche forniscono un'organizzazione formale dei principali concetti del mondo del diritto, ma alle classi ontologiche raramente corrisponde l'informazione relativa alla loro realizzazione lessicale. Sebbene sia questa una tendenza comune anche nelle ontologie formali non di dominio (Scheffczyk et al., 2006a), tuttavia una tale mancanza costituisce un aspetto particolarmente problematico nel caso della conoscenza giuridica così strettamente legata alla lingua che la esprime. È stato questo infatti uno dei motivi guida della costruzione di JurWordNet, che, grazie ai suoi collegamenti con alcune delle classi ontologiche della "Core Legal Ontology" (CLO), può essere anche visto come un'ontologia linguistica.

Coerentemente con la scelta adottata nell'ambito del progetto JurWordNet, si è qui deciso di fare riferimento alla CLO per specializzare alcuni dei STs già presenti in FrameNet o per aggiungerne di nuovi nel caso non ne fossero stati previsti.

I risultati di questo processo di specializzazione sono contenuti nella Tabella 7.14, dove per ognuno dei frames considerati adatti per descrivere le tre modalità di 'obbligo', 'permesso' e 'divieto', sono riportati i FEs oggetto di attenzione (prima colonna), l'eventuale ST presente in FrameNet (seconda colonna) e il ST proposto rappresentato da una classe ontologica della CLO (terza colonna).

[43]Nella "Core Legal Ontology" le classi qui considerate sono così definite: – **Duty**: "The obligation to do a given thing." – **Legal Subject**: "Legal Subjects (or Agents or Persons)

FE	ST in FrameNet	ST proposto (classe CLO)[43]
Frame: Being_obligated		
Duty	–	Duty
Responsible_party	–	Legal Subject
Condition	–	Condition
Frame: Being_obligatory		
Duty	–	Duty
Responsible_party	–	Legal Subject
Condition	–	Condition
Frame: Imposing_obligation		
Duty	–	Duty
Responsible_party	–	Legal Subject
Principle	–	Regulative Norm
Obligator	–	Legally Constructed Institution
Condition	–	Condition
Frame: Permitting		
State_of_affairs	–	Power
Principle	–	Power Conferring Rule
Circumstances	–	Condition
Frame: Deny_permission		
Action	–	Legal Behaviour
Protagonist	Sentient	Legal Subject
Authority	–	Legally Constructed Institution
Circumstances	–	Condition
Frame: Prohibiting		
State_of_affairs	–	Legal Behaviour
Principle	Artifact	Regulative Norm
Circumstances	–	Condition

Tabella 7.14: Le proposte di specializzazione dei STs di FrameNet con classi ontologiche della CLO.

are legally-constructed social agents, i.e. introduced by constitutive norms." – **Condition**: "A clause which makes the validity of a legal instrument or act depend on a contingency." – **Regulative Norm**: "Regulative Norms provide constraints on existing ground entities, i.e. they have situations in their scope which eventually satisfy the regulative norm (in either positive or negative sense). Regulative Norms define Behaviour Courses, and have at least one Modal Description as a proper part." – **Legally Constructed Institution**: "Legally–constructed Institutions (e.g. Ministries, Bodies, Societies, Agencies) are legal agents that perform legal acts, on behalf of powers conferred by means of power–conferring norms. They are created by constitutive norms that justify their existence and validity." – **Power**: "The fact of being capable of having rights and duties." – **Power Conferring**

L'attenzione è stata prevalentemente posta sulla caratterizzazione ontologica dei 'Core' FEs (segnalati in grassetto). Come si può vedere, per la maggior parte di essi in FrameNet non sono stati previsti STs. Per supplire a questa mancanza, ne sono stati qui suggeriti alcuni ex novo.

Diverso è il caso dei FEs Protagonist e Principle, parte (rispettivamente) del frame DENY_PERMISSION e PROHIBITING, per i quali sono stati previsti in FrameNet due STs 'generici'. La presente proposta consiste pertanto nello specializzare entrambi, chiarendo così che le istanze lessicali di questi due FEs rintracciabili in AMBnorm(Stato) sono ontologicamente caratterizzate (rispettivamente) come 'soggetti giuridici' e 'norme regolative'.

Come suggerito da Scheffczyk et al. (2006a), un tale processo di restrizione di selezione semantica di dominio ha lo scopo di fornire un ausilio *i)* alla definizione di una metodologia di annotazione semantica automatica di testi giuridici, che potrebbe essere guidata dall'informazione relativa alla realizzazione lessicale tipica di un FE, e *ii)* all'arricchimento di ontologie di dominio con informazione lessicale.

7.6.3 Specializzazioni di frames

Il caso più delicato di specializzazione di dominio riguarda l'aggiunta di nuovi frames. Come ricordato da Dolbey (2009), ci sono pro e contro nel prevedere nuovi frames oltre a quelli già presenti in FrameNet: da un lato, l'aggiunta di un frame che codifichi una situazione–tipo specifica per un determinato dominio ha il vantaggio di arricchire la risorsa generale, permettendo di rendere espliciti, in fase di annotazione, contenuti proposizionali prima non considerati; d'altro canto, tuttavia, aggiungendo un nuovo frame si corre il rischio di aumentare la complessità della rete di frames preesistente.

Sulla scia di queste riflessioni, dal momento che gli esperimenti di annotazione condotti nell'ambito di questo lavoro hanno messo in luce numerose situazioni che i frames già parte di FrameNet non erano in grado di rappresentare, si è deciso di adottare una strategia selettiva che tenesse conto della diversa natura del contenuto informativo da rendere esplicito. A questo scopo, sono state pertanto distinte le aggiunte che riguardavano casi di rappresentazione non esaustiva di realtà deontiche dai casi che avevano messo in

Rule: "A constitutive norm that confers a power to some legal role or figure." – **Legal Behaviour**: "A legal task defined by a regulative norm."

luce come nessuno dei frames presenti in FrameNet permettesse di descrivere in maniera soddisfacente gli obblighi, i permessi e i divieti regolati.

Come discusso nel Paragrafo 7.3.3, la metodologia di annotazione esplicitamente messa a punto allo scopo di rintracciare in AMBnorm(Stato) i diversi tipi di comportamenti oggetto di normazione ha infatti permesso di identificare una serie di casi un cui sarebbe necessario aggiungere nuovi frames alla rete già esistente. È questo il caso soprattutto di comportamenti che rimandano a situazioni 'extragiuridiche' evocate da LUs fattuali specialistiche relative alla materia ambientale legislata.

Sebbene in quell'occasione questo aspetto non fosse stato discusso, è tuttavia anche il caso di comportamenti che fanno riferimento a situazioni 'giuridiche' e 'extragiuridiche' generali, come dimostrano i due seguenti esempi:

a) [BEING_OBLIGATED] *Per la realizzazione delle opere previste nelle convenzioni già assentite alla data del 30 giugno 2002, ovvero rinnovate e prorogate ai sensi della legislazione vigente,* [*i concessionari* $_{Responsible\_party}$] *sono* ***tenuti*** [*ad appaltare a terzi una percentuale minima del 40 per cento dei lavori* $_{Duty}$*], [applicando le disposizioni della presente legge ad esclusione degli articoli 7, 14, 19, commi 2 e 2-bis, 27, 32, 33* $_{Condition}$].

b) [BEING_OBLIGATED] ([*Al verificarsi di un incidente rilevante,* $_{Condition}$] [*il gestore* $_{Responsible\_party}$] *è* ***tenuto*** *a* [DNI $_{Duty}$]:)[44] *c) aggiornare le informazioni fornite, qualora da indagini più approfondite emergessero nuovi elementi che modificano le precedenti informazioni o le conclusioni tratte.*

Nel primo periodo, l'istanza del FE Duty è una situazione di 'appalto', specifica del dominio giuridico, che nessuno dei frames presenti in FrameNet permette di descrivere. Nel secondo, il dovere che *il gestore* è tenuto ad adempiere, quello cioè di *aggiornare le informazioni fornite*, è relativo ad una situazione non specifica di nessun dominio ma tuttavia non descritta da nessuno dei frames presenti in FrameNet. In questo caso, è da notare

[44] In base alle scelte di segmentazione del testo in periodi descritte nel Paragrafo 3.3.1.1, questa parte del testo nella quale è istanziato il frame 'deontico' è contenuta nel periodo precedente, ma per chiarezza è stata qui riportata.

che in FrameNet esiste un frame CAUSE_CHANGE[45] il quale descrive una situazione–tipo generica 'iperonima' di quella contenuta in b), ma non in grado di rappresentare pienamente la semantica del periodo.

Nonostante la presenza di casi come questi appena considerati, coerentemente con l'oggetto del caso di studio presentato in questo capitolo, finalizzato a sperimentare i principi organizzativi di FrameNet per la rappresentazione del contenuto deontico di AMBnorm(Stato), si è scelto di discutere in quanto segue unicamente i casi in cui l'aggiunta di nuovi frames consentirebbe di rendere espliciti aspetti deontici del discorso giuridico sin'ora non considerati. Tali casi riguardano la possibilità di:

- rendere esplicito il significato di parole antonime, LUs evocatrici di situazioni–tipo che è importante rappresentare in modo distinto;
- catturare nuove prospettive di osservazione su di una situazione–tipo già esistente in FrameNet.

7.6.3.1 L'aggiunta di frames 'antonimi'

Come messo in luce nel Paragrafo 6.2.2, una delle principali caratteristiche di FrameNet è il trattamento non tradizionale della relazione di antonimia. A differenza infatti di WordNet dove due parole antonime fanno parte di due synsets diversi e sono legate da una relazione 'is–a' di antonimia, in FrameNet coppie di antonimi come ad esempio *caldo/freddo*, *amare/odiare* sono considerate evocatrici rispettivamente dei frames POSITION_ON_A_SCALE e EXPERIENCER_FOCUS. Entrambi gli elementi della coppia contribuiscono cioè a descrivere la medesima situazione–tipo, tratteggiandone due diverse polarità[46].

Sebbene una tale scelta di organizzazione del significato abbia il vantaggio di facilitare alcuni compiti di gestione dell'informazione semantica focalizzandosi sul frame piuttosto che sul materiale lessicale, tuttavia essa non consente di rendere conto di alcune distinzioni centrali in ambito giuridico.

[45] Il frame è così definito in FrameNet: "An Agent or Cause causes an Entity to change, either in its category membership or in terms of the value of an Attribute. In the former case, an Initial_category and a Final_category may be expressed, in the latter case an Initial_value and a Final_value can be specified."

[46] Come precedentemente ricordato, viene in questi casi fatta distinzione tra 'Negative' e 'Positive' LU.

La fase di annotazione semantica ha permesso di individuare un esempio significativo in questo senso. Si tratta del frame COMPLIANCE, che sulla base della definizione proposta in FrameNet[47] è evocato da LUs che descrivono uno status sia di 'conformità' sia di 'violazione' di una norma. Ciononostante, in un testo giuridico, ai fini della gestione adeguata del suo contenuto informativo, è fondamentale fare distinzione tra le due tipologie di informazione. Come ricorda infatti Wyner (2008, p. 19), "one of the distinctive characteristics of the deontic concepts is that they are violable". Di conseguenza, la possibilità di rintracciare all'interno di un corpus normativo scenari di adempimento tenendoli distinti da quelli di violazione è di fondamentale importanza.

Per questo motivo, si suggerisce qui di aggiungere un nuovo frame VIOLATION per descrivere in modo separato la situazione–tipo evocata dagli antonimi delle LUs evocatrici del frame COMPLIANCE. Il nuovo frame potrebbe essere legato da una relazione 'frame–to–frame' di tipo Inheritance al frame preesistente, ereditandone pertanto i FEs ad eccezione tuttavia del 'Core' FE Norm che si potrebbe in questo caso chiamare Violated_norm.

Come esemplificato nei due seguenti periodi annotati in AMBnorm(Stato), una tale modifica consentirebbe di rendere esplicito il fatto che le norme a cui si fa riferimento non sono quelle alle quali il comportamento di un soggetto giuridico si deve attenere, ma sono quelle espressione delle regole di comportamento violate:

a) [VIOLATION] *Le* ***violazioni*** [*delle disposizioni di cui all'articolo 12 in materia di vendita a distanza* $_{Violated\_norm}$] *sono punite con la sanzione amministrativa pecuniaria da euro mille ad euro settemilacinquecento.*

b) [VIOLATION] *Il fabbricante o il mandatario che immette in commercio o mette in servizio macchine ed attrezzature di cui all'allegato I, parte c), [in* $_{Supp}$] ***violazione*** [*alle disposizioni di cui all'articolo 11, comma 2* $_{Violated\_norm}$]*, è punito, fuori dai casi in cui la violazione costituisce reato, con la sanzione amministrativa pecuniaria del pagamento di una somma da euro 1000 a euro 50000.*

L'importanza di annotare in modo specifico l'informazione relativa alla 'violazione' di una norma di comportamento, tenendola distinta da quella relativa alla sua 'ottemperanza', è ancor più evidente assumendo una prospettiva di annotazione a testo continuo. Questa modalità di annotazione

[47] Vedi Paragrafo 7.1.

permette infatti di chiarire le conseguenze della violazione delle *disposizioni*, che se non rispettate saranno punite. Come mostrano le seguenti annotazioni aggiunte ai periodi a) e b) precedenti, una tale informazione è rappresentata dal frame REWARDS_AND_PUNISHMENTS (previsto in FrameNet):

a.1) [REWARDS_AND_PUNISHMENTS] [*Le violazioni delle disposizioni di cui all'articolo 12 in materia di vendita a distanza* $_{Reason}$] *sono* ***punite*** [*con la sanzione amministrativa pecuniaria da euro mille ad euro settemilacinquecento* $_{Response\_action}$].

b.1) [REWARDS_AND_PUNISHMENTS] [*Il fabbricante o il mandatario che immette in commercio o mette in servizio macchine ed attrezzature di cui all'allegato I, parte c), in violazione alle disposizioni di cui all'articolo 11, comma 2* $_{Evaluee}$], *è* ***punito***, *fuori dai casi in cui la violazione costituisce reato,* [*con la sanzione amministrativa pecuniaria del pagamento di una somma da euro 1000 a euro 50000* $_{Response\_action}$].

7.6.3.2 Aggiunta di nuove prospettive di osservazione

Un esempio degno d'interessante che dimostra come l'aggiunta di uno o più frames consentirebbe una descrizione più dettagliata del contenuto informativo di testi giuridici riguarda la specializzazione del frame PERMITTING.

Come si può vedere nella precedente Figura 7.1, il frame è legato da una relazione di Inheritance al frame PROHIBITING, a sua volta legato da una relazione di tipo Using al frame LAW. Questa rete di relazioni 'frame–to–frame' permette di mettere in luce come i due frames descrivano due situazioni–tipo parallele nelle quali un determinato stato di cose è 'permesso' o 'proibito' da un principio regolativo, da un insieme di norme giuridiche. In entrambi sono infatti previsti due soli 'Core' FEs: Principle e State of affairs.

Inoltre, nelle definizioni di entrambi è specificato che il principio regolativo non è un'autorità che permette o proibisce a qualcuno di fare qualcosa[48]. Il caso in cui un'agente si rivolge ad un soggetto per negare o concedere un permesso è infatti descritto dai frames DENY_PERMISSION e GRANT_PERMISSION. La relazione di tipo Using che lega questi due frames al frame COMMUNICATION chiarisce che si tratta di situazioni–tipo non giuridicamente caratterizzate, ma di semplici atti di comunicazione.

[48]Vedi la definizione del frame PROHIBITING e PERMITTING al Paragrafo 7.1.

Tuttavia, mentre il frame DENY_PERMISSION non prevede restrizioni in questo senso, la definizione della situazione–tipo GRANT_PERMISSION prevede che "this frame does not include situations where there is a state of permission granted by authority or rule of law"[49]. Di conseguenza, in FrameNet viene offerta una rappresentazione non esaustiva del concetto di 'permesso'. In particolare, non è prevista la presenza di una situazione–tipo nella quale *i)* sia un'agente giuridico (un'autorità) a conferire un 'potere' e nella quale *ii)* il 'potere' sia destinato ad un determinato soggetto giuridico. Al contrario, entrambe le informazioni sono centrali per il dominio giuridico, nel quale gli enunciati deontici devo sempre essere indirizzati ad un destinatario.

La soluzione proposta in questo lavoro è dunque quella di aggiungere un nuovo frame GRANT_LEGAL_PERMISSION, che specializza il frame preesistente GRANT_PERMISSION. Ciò permetterebbe di descrivere anche per lo scenario di 'permesso' una situazione–tipo analoga a quella descritta dal frame DENY_PERMISSION. Il nuovo frame potrebbe dunque essere legato da una relazione di tipo Inheritance al frame GRANT_PERMISSION e potrebbe contenere i tre seguenti 'Core' FEs, ereditandoli (e specializzandoli) da quelli del frame già esistente in FrameNet[50]:

- 'Legal grantor' < 'Grantor',
- 'Grantee' < 'Grantee',
- 'Permitted_action' < 'Action'.

Gli esempi che seguono dimostrano come questo nuovo frame qui suggerito permetterebbe di assumere una diversa prospettiva di osservazione sul 'permesso', rendendo pienamente esplicita l'informazione contenuta nel periodo[51]:

[49] A livello infatti delle relazioni tra i singoli FEs, i due frames differiscono tra loro nel rapporto con il frame COMMUNICATION. Mentre i due frames DENY_PERMISSION e COMMUNICATION sono legati da una relazione tra i 'Core' FEs Authority e Protagonist del primo con i 'Core' FEs Communicator e Topic del secondo, il frame GRANT_PERMISSION è legato al frame COMMUNICATION dalla relazione che intercorre tra i suoi 'Core' FEs Grantor e Grantee rispettivamente con i 'Core' FEs Communicator e Addressee del frame COMMUNICATION.

[50] A sinistra del segno < sono elencati i FEs del nuovo frame GRANT_LEGAL_PERMISSION, a destra i corrispondenti FEs del preesistente frame GRANT_PERMISSION.

[51] La LU evocatrice del frame GRANT_LEGAL_PERMISSION è segnalata in grassetto.

a) [GRANT_LEGAL_PERMISSION] [*In sede di revisione catastale* $_{Circumstances}$], [*è data* $_{Supp}$] ***facoltà*** [*agli enti locali* $_{Grantee}$] [*, con proprio provvedimento,* $_{Means}$] [*di disporre l'accorpamento al demanio stradale delle porzioni di terreno utilizzate ad uso pubblico, ininterrottamente da oltre venti anni, previa acquisizione del consenso da parte degli attuali proprietari* $_{Permitted\_action}$]. [CNI $_{Legal\_grantor}$]

b) [GRANT_LEGAL_PERMISSION] [*Il Ministero della sanità, per quanto riguarda gli aspetti ambientali d'intesa con il Ministero dell'ambiente* $_{Legal\_grantor}$], ***autorizza*** [*ai sensi del presente decreto* $_{Circumstances}$] [*l'immissione sul mercato e l'utilizzazzione nel territorio italiano di un biocida* $_{Permitted\_action}$].

7.7 Considerazioni conclusive

Le discussioni condotte in questo capitolo hanno permesso di mettere in luce i vantaggi e i limiti dell'adozione dei principi di organizzazione del significato di FrameNet nell'annotazione semantica di testi normativi, finalizzata a rendere esplicito il modo in cui l'informazione deontica è in essi organizzata.

In primo luogo, la rassegna dei frames presenti in FrameNet ha rivelato come le tre principali modalità deontiche di 'obbligo', 'permesso' e 'divieto' siano rappresentate da almeno un frame. Sebbene infatti siano state proposte nel Paragrafo 7.6 alcune specializzazioni per descrivere in modo più esaustivo aspetti non pienamente considerati, tuttavia le situazioni–tipo previste nella rete di frames in FrameNet permettono di rendere esplicita la qualifica deontica di comportamenti imposti, permessi e vietati nel corpus di testi normativi preso in considerazione in questo caso di studio.

È stato inoltre possibile verificare come FrameNet consenta di descrivere la modalità anankastica di un comportamento richiesto grazie alla presenza di un frame REQUIRED_EVENT. La differenza rispetto alla modalità deontica di un comportamento obbligatoriamente imposto (modalità rappresentata dal frame OBLIGATION_SCENARIO) è in particolare espressa a livello dei singoli FEs dei due frames.

Dal momento che, tra i tre scenari deontici considerati, quello di 'obbligo' si è rivelato quello meglio descrivibile con gli elementi di rappresentazione semantica offerti da FrameNet, è su questa modalità deontica che si è concentrato gran parte del caso di studio condotto. L'annotazione delle istanze di

scenari di obbligo presenti nel corpus AMBnorm(Stato) ha infatti permesso di dimostrare come FrameNet sia affidabile come modello per:

- raccogliere informazioni lessicografiche relative alle proprietà combinatorie a livello sintattico e semantico di termini evocatori di una situazione deontica. Il fatto di assumere come punto di partenza per la descrizione del significato lessicale il **testo** e non la competenza a priori del lessicografo ha permesso di focalizzare l'attenzione in particolare su come sia possibile adottare FrameNet come modello per la costruzione di un lessico giuridico non già basato sull'idea che il significato di una parola sia qualcosa di "intrinsecamente e definitivamente legato ad essa"[52], ma sull'idea che esso sia descrivibile a partire dalle sue occorrenze d'uso;

- rendere esplicito il contenuto proposizionale dei periodi annotati a partire dalla loro struttura sintattica. In questo senso, gli esperimenti di annotazione semantica condotti hanno dimostrato come il principio di FrameNet di separare "the notion of the conceptual underpinnings of a concept from the precise way in which the words anchored in them get used" (Fillmore e Atkins, 1992) permetta di mettere in luce come alcuni dei più distintivi comportamenti sintattici di AMBnorm(Stato), individuati in fase di monitoraggio linguistico[53], influenzino l'organizzazione del contenuto semantico nel testo;

- rappresentare in modo formale e organizzato il contenuto informativo dei periodi annotati e organizzarlo sulla base dei principi organizzativi del significato di FrameNet. Ciò ha in particolare permesso di dimostrare come *i)* il modello 'frame–based' di organizzazione della conoscenza adottato in FrameNet consenta di rappresentare aspetti del 'concetto giuridico fondamentale' **obbligo** non considerati nelle ontologie giuridiche e come *ii)* il livello di organizzazione a livello sintagmatico del significato aiuti a catturare il contesto d'uso dei termini espressione dei concetti, a differenza di quanto avviene sino ad oggi nelle ontologie giuridiche.

[52] La definizione è di Scarpelli (1976b).
[53] Vedi Paragrafo 4.

I risultati degli esperimenti di annotazione condotti hanno inoltre permesso di mettere in luce i vantaggi delle novità introdotte in questo studio rispetto al modello originario per quanto riguarda in particolare:

- la scelta di assumere come punto di partenza dell'annotazione semantica il risultato dell'annotazione sintattica a dipendenze realizzata in modo automatico. Ciò ha permesso infatti di
  - guidare l'annotazione semantica manuale, riducendo i casi di incoerenza dovuti all'erroneo riconoscimento di istanze di FEs non legate (dipendenti) alla LU evocatrice,
  - rendere esplicito il modo in cui la semantica è veicolata da sezioni specifiche dell'intero albero sintattico di un periodo giuridico, superando in questo modo il ben noto limite delle annotazioni di FrameNet non legate a strutture sintattiche globali del periodo;
- la selezione semi–automatica di LUs poliremathiche (come ad esempio le costruzioni a verbo supporto) che ha dimostrato
  - come alcune delle caratteristiche morfosintattiche di AMBnorm(Stato), rilevate in fase di monitoraggio del corpus a questo livello di descrizione linguistica[54], abbiano conseguenze sulla distribuzione del carico semantico portato dalle diverse categorie morfosintattiche nel testo,
  - come il carattere 'formulaico' della lingua del diritto permetta di individuare con successo sequenze di singole unità lessicali che costituiscono vere e proprie "formule", più o meno fisse, dotate di un "significato finito e specialistico sulla base della loro forza associativa"[55];
- la definizione di un'innovativa modalità di annotazione finalizzata a rendere espliciti il modo in cui la realtà espressione del mondo del diritto si intrecci con quella del mondo dei fatti regolati dal diritto Ciò ha

[54] Vedi Paragrafo 4.2.2.

[55] La citazione è di Eklund-Braconi (2000).

– contribuito alla completa rappresentazione del contenuto informativo dei periodi annotati, grazie all'annotazione non solo della modalità deontica ma anche della situazione–tipo imposta, permessa o vietata;

– suggerito una possibile soluzione alla "epistemological promiscuity", per dirla con le parole di Breuker e Hoekstra (2004), di cui soffrono molti sistemi di organizzazione della conoscenza giuridica, sistemi nei quali cioè l'informazione relativa ai concetti espressione della realtà giuridica è indiscriminatamente mischiata con quella relativa alla realtà extragiuridica regolata.

Infine, i limiti maggiori dell'adozione di FrameNet come modello per l'annotazione semantica di testi giuridici sono tutti riconducibili al fatto che esso è stato pensato e sviluppato per la rappresentazione del significato di testi giornalistici assunti come rappresentativi della lingua comune. È questo il motivo per cui in questo studio sono state proposte una serie di specializzazioni della risorsa originaria, restringendo però il campo ai casi che riguardano modifiche legate alla rappresentazione esaustiva della sola informazione deontica.

In aggiunta, gli esperimenti di annotazione hanno permesso di mettere in luce una serie di ulteriori specializzazioni che potrebbero essere apportate. Tra le altre, è stato fatto notare come non sempre i frames presenti in FrameNet consentano, ad esempio, di rendere esplicite situazioni–tipo specifiche del dominio ambientale, della realtà tecnico–giuridica e di quella comune regolate. È questo uno dei motivi per cui è importante ricordare come il caso di studio qui descritto non miri ad essere esaustivo, ma costituisca piuttosto un primo esempio di come FrameNet possa essere applicato con successo come modello di annotazione semantica di testi giuridici.

Capitolo 8

Conclusioni

Nell'introduzione di questo studio erano state poste una serie di domande circa il dibattuto rapporto tra analisi linguistica di testi giuridici e accesso al loro contenuto, domande alle quali ci si riproponeva di suggerire alcune risposte utilizzando metodi e strumenti linguistico–computazionali.

Nel tracciare ora le considerazioni conclusive di questo lavoro è dunque intenzione ripercorrere gli interrogativi di partenza ed esporre le soluzioni proposte rispetto ai diversi aspetti di indagine presi in considerazione.

Aspetti metodologici

Il primo risultato di questo lavoro è di tipo metodologico. Nell'introduzione era stato fatto notare che il principale obiettivo dell'intero studio era quello di trovare una metodologia di analisi che permettesse di rendere esplicite le relazioni tra la struttura sintattico–grammaticale di un testo giuridico e il modo in cui vi è organizzato il contenuto semantico–informativo. Il proposito era quello di mettere a punto una metodologia di indagine del testo che permettesse di rendere effettivo l'invito a "porsi questioni linguistiche in stretta connessione con questioni giuridiche"[1] che Bice Mortara Garavelli rivolge al linguista, nonché l'indicazione dei giuristi relativamente al fatto che "i problemi di significato degli enunciati giuridici possono essere affrontati solo risolvendone i problemi sintattici"[2].

La soluzione adottata è consistita nell'utilizzare strumenti di annotazione linguistica automatica del testo come punto di partenza per accedere in

[1] Garavelli (2001, p. 34).
[2] Jori e Pintore (1995, p. 209).

maniera incrementale al contenuto informativo e darne una strutturazione formale. In questo senso dunque, tra gli altri vantaggi, anche dal punto di vista metodologico la scelta di FrameNet come modello di organizzazione e rappresentazione del significato, nonché come modello di annotazione semantica del testo, è stata fondamentale. Come chiarito dai suoi ideatori stessi "the job of FrameNet is to document from attested instances of contemporary English the manner in which frame elements (for given words in given meanings) are grammatically instantiated in English sentences and to organize and exhibit the results of such findings in a systematic way"[3].

Essendo finalizzato a ciò, FrameNet si è pertanto rivelato un punto di partenza ottimale per definire una metodologia di indagine che permette di condurre uno studio completo dei testi giuridici. Elemento chiave di FrameNet è infatti il principio di basare il processo di annotazione semantica sul livello di annotazione sintattica. Ciò consente di mettere chiaramente in luce il rapporto tra organizzazione sintattica e semantica del materiale informativo rilevante in un testo.

Aspetti di indagine linguistica

Presentando alcuni degli aspetti di indagine sui quali i linguisti hanno concentrato le loro attività di ricerca, era stata posta particolare attenzione sull'osservazione di Michele Cortelazzo riguardo al carattere "multiforme e complesso" della lingua del diritto. Per Cortelazzo (1997) tale carattere è principalmente riconducibile sia alla varietà di tipologie di testi nei quali la lingua si instanzia, sia ai suoi stretti e biunivoci rapporti con la lingua comune e i linguaggi tecnico–specialistici, cioè alle sue "articolazioni orizzontali (per sottosettori del diritto)" e "verticali (con distinzioni fra espressioni puramente tecniche ed espressioni di uso comune)"[4].

Ponendosi come obiettivo quello di trovare una strategia di indagine linguistica in grado di affrontare il tema della complessità della lingua del diritto, l'approccio messo a punto e descritto nel Capitolo 4 si è rilevato affidabile per condurre uno studio di testi giuridici in grado di suggerire alcune risposte ai due aspetti problematici individuati da Cortelazzo.

La metodologia comparativa di monitoraggio linguistico adottata ha permesso infatti, da un lato, di descrivere le caratteristiche lessicali, morfosintattiche e sintattiche del corpus di testi giuridici preso in esame confrontando

[3]Fillmore e Baker (2001).
[4]Vedi Cortelazzo (1997).

il modo in cui alcuni significativi tratti linguistici si distribuiscono in questi testi e in testi giornalistici assunti come rappresentativi della lingua comune. Dall'altro, il punto di vista 'interno' allo studio della lingua del diritto ha permesso di mettere in luce come i vari tratti linguistici si distribuiscano diversamente nei diversi tipi di testi giuridici esaminati.

Inoltre, la scelta di prendere come termini di confronto due corpora rappresentativi di due diverse varietà di prosa giornalistica è stata espressamente finalizzata a indagare più nel dettaglio le similarità e differenze della lingua di testi giuridici rispetto alla lingua comune. Questo ha infatti permesso di verificare empiricamente fino a che punto la lingua dei testi giuridici si differenzi, da un lato, da quella usata in testi comuni che dovrebbero essere leggibili ad un ampio pubblico di lettori, come gli articoli del quotidiano "La Repubblica", e, dall'altro, da quella pensata per essere estremamente semplice, come quella dei testi parte del mensile "Due Parole", volutamente scritti per essere letti e compresi da persone con un basso livello di alfabetizzazione o con ridotte capacità cognitive.

È stato in questo modo possibile dimostrare empiricamente come la domanda posta da Fiorelli (2008) a proposito del posto occupato dalla lingua del diritto in una possibile classificazione di linguaggi specialistici basata sui rapporti con la lingua comune non possa che avere una risposta quanto mai articolata. Essa non interessa infatti solo l'aspetto lessicale preso in considerazione da Fiorelli, ma coinvolge una nutrita serie di tratti linguistici relativi alla distribuzione delle categorie morfosintattiche, alla struttura sintattica del periodo, ai valori di ricchezza lessicale, ecc..., tratti che interagiscono in modo complesso tra di loro permettendo di definire chiaramente le differenze tra testi giuridici e testi giornalistici e tra diverse varietà di atti giuridici.

Infine, coerentemente con il generale approccio metodologico seguito, l'individuazione delle principali caratteristiche soprattutto sintattiche dei testi giuridici ha costituito il punto di partenza per analizzare come esse si intreccino con caratteristiche relative ai modi di organizzazione nel testo del contenuto semantico. Le possibilità espressive offerte da FrameNet hanno permesso infatti di rendere esplicito come gli elementi conoscitivi presenti in un periodo si istanzino in particolari strutture sintattiche. È stato così possibile suggerire una strategia di indagine metodologicamente diversa da quella proposta da Rovere (2005) ma ugualmente finalizzata a mettere in luce come la distribuzione di caratteristiche morfosintattiche e sintattiche rintracciate in corpora di testi giuridici siano indicative e rilevanti per un loro studio a livello semantico.

Aspetti di trattamento automatico della lingua del diritto

Come ricordato nel Paragrafo 2.3.2, uno degli aspetti storicamente al centro del dibattito in materia di AI&Law riguarda la difficoltà di gestire in modo automatico il contenuto di testi giuridici a causa della lingua "convoluted and unnatural"[5] nella quale sono scritti. In una tale ottica applicativa, la questione è legata alla diminuzione della precisione degli strumenti di annotazione linguistica automatica nell'analisi di testi giuridici.

Questione centrale per ogni successivo compito di gestione della conoscenza, in questo studio essa è stata messa in relazione principalmente con la fase di annotazione semantica del testo. In un'ottica di futura automatizzazione del processo, infatti, individuare con buon livelllo di affidabilità le strutture linguistiche (sintattiche) nelle quali si istanziano gli elementi semanticamente rilevanti nel testo è fondamentale per replicare in modo automatico le annotazioni semantiche.

Tema per lo più trascurato sia dai linguisti computazionali sia dalla comunità di ricerca in AI&Law, esso è tuttavia al centro di alcuni dei più recenti lavori di chi sviluppa applicazioni semantiche basate sull'uso di strumenti di Trattamento Automatico del Linguaggio[6]. Ciononostante, per quanto riguarda la lingua italiana in particolare, mancano sino ad oggi studi dedicati a valutare dal punto di vista quantitativo l'impatto che la lingua del diritto ha sull'accuratezza dell'annotazione linguistica automatica, soprattutto in un approccio 'data–driven' all'elaborazione automatica del testo.

Uno dei contributi più innovativi dello studio qui condotto consiste dunque nell'aver *i)* quantificato l'accuratezza dell'annotazione sintattica a dipendenze realizzata da strumenti 'data–driven' attraverso il confronto con i risultati delle analisi dei testi giornalistici, rappresentativi della lingua comune, sui quali gli strumenti sono stati addestrati; *ii)* individuato le principali strutture sintattiche responsabili della diminuzione dell'accuratezza d'analisi; *iii)* proposto, di conseguenza, alcune soluzioni di annotazione alternative, specializzando i criteri di annotazione seguiti per l'analisi dei testi rappresentativi della lingua comune; *iv)* creato (sulla base di quanto esposto nei punti precedenti) un corpus di testi giuridici (legislativi) annotati in modo manuale fino al livello sintattico, messo a disposizione della comunità di linguistica computazionale per la valutazione di strumenti di Trattamento Automatico

[5] McCarty (2009).
[6] Vedi Paragrafo 2.3.2.1.

del Linguaggio sviluppati per l'annotazione di corpora di lingua comune e usati nell'analisi di testi giuridici[7].

Aspetti di accesso al contenuto testuale

Come discusso nel Paragrafo 5.1, una ben nota peculiarità dei testi giuridici, e di quelli normativo–amministrativi in particolare, è quella di essere caratterizzati da un "complesso intreccio di realtà giuridiche ed extragiuridiche" che si riflette nel loro lessico, per dirla con le parole di Belvedere (1994a).

Questione al centro del dibattito teorico degli studi condotti da linguisti e filosofi del diritto, da un punto di vista applicativo essa è fonte di difficoltà soprattutto *i)* in un'ottica di rappresentazione formale della conoscenza giuridica finalizzata, ad esempio alla costruzione di ontologie giuridiche nelle quali non siano indiscriminatamente mischiati il livello di descrizione del mondo dei fatti oggetto di regolamentazione (espressione della realtà extragiuridica) e la descrizione delle primitive di conoscenza (i concetti giuridici) rappresentative della realtà giuridica[8]; *ii)* per approcci basati su di un'esplicita attenzione ai termini come principale via d'accesso al contenuto di testi giuridici, quali l'estrazione di terminologia rilevante da corpora testuali.

La soluzione suggerita in questo studio ha riguardato i due passi nei quali si articola l'approccio all'accesso al contenuto di testi giuridici qui messo a punto e descritto nel Capitolo 5. Esso consiste infatti *i)* nella definizione di una metodologia di estrazione automatica di terminologia che permette di discriminare le diverse tipologie di lessico presenti in un testo giuridico[9] e *ii)* nella definizione di un'innovativa modalità di annotazione semantica che, specializzando le modalità previste nel progetto FrameNet, consente di annotare in modo separato la componente fattuale da quella deontico–giuridica contenuta in periodi giuridici[10].

Aspetti di rappresentazione del significato e della conoscenza

Uno degli aspetti non pienamente soddisfacenti in materia di rappresentazione della conoscenza contenuta in corpora di testi giuridici riguarda, come

[7]Si ricorda qui come tale corpus sia attualmente usato nell'ambito del "Domain Adaptation for Dependency Parsing Task" dell'edizione 2011 di Evalita. Vedi a questo proposito il Paragrafo 3.5.

[8]Vedi il dibattito riportato da Breuker e Hoekstra (2004).

[9]Vedi Paragrafo 5.2.

[10]Vedi Paragrafo 7.3.3.

discusso nel Paragrafo 6.5.2, la mancata attenzione al contesto nel quale i termini, i primi oggetti linguistici istanze degli oggetti di conoscenza, si collocano. Secondo la motivazione esposta da Breuker (2009), ciò è dovuto al fatto che, sino ad oggi ci si è concentrati solo sull'aspetto paradigmatico di rappresentazione del significato, trascurando quello sintagmatico.

La metodologia di annotazione semantica basata sul modello FrameNet è la risposta suggerita in questo studio ad una tale questione. Essa si ispira, da un lato, agli studi sul lessico del diritto e di quelli in semiotica giuridica condotti nell'alveo della scuola di filosofia analitica del diritto[11] e, dall'altro, all'approccio 'frame–based' alla rappresentazione della conoscenza adottato da van Kralingen (1997) negli anni '90 nella costruzione della "Frame–Based Ontology of Law".

Espressamente finalizzata a rappresentare le proprietà semantico–combinatorie delle parole a partire dalle strutture sintattiche nelle qualli esse ricorrono nel testo, una tale metodologia si configura come un modello complementare al modello WordNet sino ad oggi utilizzato per l'organizzazione della conoscenza semantico–lessicale giuridica, modello basato su principi di organizzazione paradigmatica del significato.

In questo senso, i vari esperimenti di annotazione discussi nel Capitolo 7 hanno permesso di mettere in evidenza come l'utilizzo dei principi di organizzazione del significato e di annotazione semantica di FrameNet, in unione alle novità e alle specializzazione introdotte in questo studio, lo rendano un modello particolarmente espressivo per aspetti di rappresentazione del significato e della conoscenza in ambito giuridico. I principi di rappresentazione del materiale semantico–lessicale su cui si basa FrameNet permettono infatti, da un lato, di rendere noto non solo "what we know about terms", ma anche "what terms mean in a particular context (domain, document, phrase, ...)", per dirla con le parole di Breuker (2009); dall'altro, essi permettono di rendere espliciti i ruoli che i vari elementi conoscitivi svolgono in un testo, contribuendo così a descrivere in modo esaustivo le diverse situazioni in esso contenute.

Sviluppi futuri

Nel concludere questo lavoro è intenzione sottolineare come i risultati ottenuti costituiscono il punto d'arrivo di un percorso di ricerca finalizzato a delineare

[11] Vedi Paragrafo 2.2.

e sperimentare una metodologia di studio del rapporto tra analisi linguistica di testi giuridici e accesso al loro contenuto che fosse innovativa rispetto agli studi condotti sino ad oggi e che permettesse di mettere in luce le potenzialità dell'uso di strumenti di Trattamento Automatico del Linguaggio ancora non pienamente esplorate in quest'ambito di studi.

D'altro canto i traguardi raggiunti rappresentano il punto di partenza di una serie di sviluppi futuri. L'intento di questa ultima parte è dunque quello di tratteggiare le direzioni di ricerca che questo studio ha aperto.

Dal punto di vista linguistico, la metodologia di monitoraggio delle caratteristiche morfosintattiche, sintattiche e lessicali descritta nel Capitolo 4 può essere *i)* estesa ad altri tipi di testi giuridici, oltre a quelli normativi e amministrativi qui presi in esame, come ad esempio le sentenze; *ii)* condotta su più ampia scala, su corpora di più ampie dimensioni e *iii)* realizzata ampliando l'insieme e la tipologia di tratti linguistici, oltre a quelli considerati in questo studio. Nel discutere i singoli tratti monitorati, era stato infatti fatto notare come alcuni di essi meritassero analisi più approfondite e come altri fossero il risultato di annotazioni linguistiche automatiche al momento non sempre sufficientemente affidabili.

Una tale metodologia potrà inoltre essere ulteriormente estesa allo scopo di sviluppare uno strumento a supporto delle attività di verifica della redazione 'chiara, semplice e comprensibile' di un atto normativo–amministrativo e di un indicatore del livello di leggibilità di testi giuridici basato sul monitoraggio linguistico[12].

Con l'intento di rendere effettive queste future direzioni di ricerca, la definizione di una strategia di adattamento di strumenti di annotazione linguistica automatica e in particolare di strumenti di annotazione sintattica a dipendenze è uno dei principali e naturali sviluppi di questo lavoro. Essa consentirebbe infatti di sviluppare strumenti più affidabili come punto di partenza per analisi linguistiche più affidabili e raffinate.

Le indagini condotte nel Capitolo 3 aprono la strada proprio a questo futuro scenario applicativo. Il corpus di testi normativi annotato in modo manuale fino al livello sintattico costruito potrà essere utilizzato come punto di partenza per addestrare un parser ('data–driven') basato su un algoritmo

[12]Si ricorda qui come questa linea di ricerca sia già avviata nell'ambito della partecipazione alle attività dell'"Osservatorio per il recepimento e l'attuazione della 'Guida per la redazione degli atti amministrativo"' insediatosi il 1 aprile 2011. Vedi a questo proposito il Paragrafo 4.3.2.

di apprendimento automatico. Quanto ci si aspetta è che questo consenta di aumentare il livello di precisione delle analisi realizzate.

Similmente, in linea con gli usi che sino ad oggi sono stati fatti di FrameNet[13], l'insieme di periodi semanticamente annotati in modo manuale può essere utilizzato in compiti di annotazione semantica automatica come il 'Semantic Role Labeling' per addestrare un cosiddetto 'parser semantico'. In tale prospettiva, la scelta di basare le annotazioni semantiche sul livello di annotazione sintattica automatica è centrale. Essa permette infatti di replicare in modo automatico le annotazioni semantiche legandole al riconoscimento della struttura sintattica sottostante, riducendo i casi di ambiguità d'annotazione. Come discusso nel Paragrafo 7.2, tale scelta pone questo lavoro tra quelli finalizzati a sviluppare risorse testuali semanticamente annotate a partire dalla struttura sintattica globale dei periodi annotati.

Inoltre, l'ampliamento della collezione di periodi semanticamente annotati sul modello FrameNet è tra i possibili sviluppi. Questo consentirebbe, da un lato, di allargare il corpus di addestramento in vista di un compito di annotazione semantica automatica, rendendo in questo modo statisticamente più affidabili le annotazioni di un futuro 'parser semantico'; dall'altro, permetterebbe di condurre studi su più ampia scala del rapporto tra realizzazione sintattica e contenuto semantico–lessicale in testi giuridici. Ammesso dai principi stessi di organizzazione e rappresentazione del significato su cui si basa FrameNet, un tale studio è stato parzialmente condotto nel Paragrafo 7.4, ma meriterebbe analisi più approfondite realizzabili solo disponendo di una maggiore quantità di periodi semanticamente annotati.

Infine, è qui importante mettere in evidenza come la metodologia di annotazione semantica basata sul modello FrameNet esemplificata nel caso di studio riportato nel Capitolo 7 apra la strada a future applicazioni. In primo luogo, essa meriterebbe di essere estensa ad un corpus di dimensioni maggiori e ulteriormente specializzata focalizzandosi su nuovi aspetti di rappresentazione del significato e della conoscenza del dominio giuridico che al momento non è stato possibile tenere in considerazione.

In questo modo essa potrebbe essere utilizzata per costruire un lessico giuridico *i)* basato su evidenza testuale e *ii)* con informazione relativa al contesto sintattico nel quale i lemmi occorrono e alla situazione–tipo (al frame) a cui rimandano. Un lessico fondato su questi principi organizzativi, da un lato, troverebbe i suoi presupposti teorici in quanto fatto osservare dai

[13]Vedi Paragrafo 6.1.3.

giuristi della scuola di filosofia analitica del diritto per i quali "il significato di una parola non è qualcosa che sia intrinsecamente e definitivamente legato ad essa"[14], ma è determinato dalle regole d'uso stabilite in un universo concettuale condiviso. Dall'altro, realizzerebbe il 'sogno' di Breuker (2009) di poter disporre di risorse lessicali e ontologiche nelle quali il significato dei termini espressione dei principali concetti sia definito sulla base del contesto in cui essi concretamente occorrono e non sulla base di astratte conoscenze a priori. In tal senso un lessico di questo tipo sarebbe dunque complementare a JurWordNet.

Una futura specializzazione di dominio dei Semantic Types (STs) permetterebbe inoltre di aggiungere nuovi possibili collegamenti con classi di ontologie giuridiche già esistenti, ampliando quelli suggeriti nel caso di studio[15]. Ciò contribuirebbe a portare a compimento uno dei propositi che hanno mosso la definizione della metodologia di annotazione semantica messa a punto in questo lavoro: quella di dimostrare come i principi di organizzazione della conoscenza di FrameNet siano complementari a quelli su cui si basano le ontologie giuridiche e come i primi possano unirsi ai secondi in un fruttuoso scambio reciproco. L'aggiunta di nuovi STs di dominio, infatti, arricchirebbe *i)* le ontologie di dominio con informazione lessicale e *i)* le annotazioni basate sul modello FrameNet con informazione ontologica di dominio.

[14]Scarpelli (1976b).
[15]Vedi Paragrafo 7.6.2.

Appendice I

In questa prima appendice sono riportati gli schemi di annotazione morfosintattica e sintattica, sviluppati dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa nell'ambito del progetto TANL ("Text Analytics and Natural Language processing")[16] e utilizzati nell'annotazione della treebank ISST–TANL.

Lo schema di annotazione morfosintattica

Lo schema di annotazione morfosintattica riportato nella Tabella 8.1, conforme allo standard EAGLES ("Expert Advisory Group for Language Engineering Standards")[17], comprende 14 categorie 'generali' (prima colonna) e 37 sottocategorie (seconda colonna).

CPoS	**FPoS**	**Descrizione**	**Esempi**	**Contesti d'uso**
A	A	Aggettivo	*bello, buono, pauroso, ottimo*	una **bella** passeggiata un **ottimo** attaccante una persona **paurosa**
	AP	Aggettivo possessivo	*mio, tuo, nostro, loro*	a **mio** parere il **tuo** libro
B	B	Avverbio	*bene, fortemente, malissimo, domani*	arrivo **domani** sto **bene**
	BN	Avverbio di negazione	*non*	**non** sto bene
C	CC	Congiunzione coordinativa	*e, o, ma, ovvero*	i libri **e** i quaderni vengo **ma** non rimango
	CS	Congiunzione subordinativa	*mentre, quando*	**quando** ho finito vengo **mentre** scrivevo ho finito l'inchiostro

[16] http://medialab.di.unipi.it/wiki/SemaWiki
[17] http://www.ilc.cnr.it/EAGLES/home.html

CPoS	FPoS	Descrizione	Esempi	Contesti d'uso
D	DE	Determinante esclamativo	*che, quale, quanto*	**che** disastro! **quale** catastrofe!
	DI	Determinanate indefinito	*alcuno, certo, tale, parecchio, qualsiasi*	**alcune** telefonate **parecchi** giornali **qualsiasi** persona
	DQ	Determinante interrogativo	*che, quale, quanto*	**che** cosa **quanta** strada **quale** formazione
	DR	Determinate relativo	*cui, quale*	i **cui** libri
	DD	Determinante dimostrativo	*questo, codesto, quello*	**questo** denaro **quella** famiglia
E	E	Preposizione	*di, a, da, in, su, attraverso, verso*	**a** casa **del** poeta **verso** sera
	EA	Preposizione articolata	*del, alla, dei, nelle*	**nella** casa il prezzo **del** pane
F	FB	Punteggiatura bilanciata	( ) “ ” ‘ ’ - -	il gatto – che conoscete –
	FC	Punteggiatura di fine frase	*, ;*	ha detto **:** Vieni!
	FF	Virgola, trattino	*,*	mele**,** pere e banane due**-**trecento persone
	FS	Punteggiatura di fine periodo	*. ? !*	mele, pere e banane**.** cosa vuoi**?**
I	I	Interiezione	*ahim, beh, ecco, grazie*	**Beh**, che vuoi?
N	N	Numero cardinale	*uno, due, cento, mille, 28, 2000*	**due** partite **28** anni
	NO	Numero ordinale	*primo, secondo, centesimo*	**secondo** posto
P	PD	Pronome dimostrativo	*questo, quello, costui*	**quello** di Roma **costui** uccide
	PE	Pronome personale	*egli, lui, esso* *noialtri, voialtri, essi* *io, me, tu, te*	**io** parto **lo** mangio
	PI	Pronome indefinito	*chiunque, ognuno, molto*	**chiunque** venga i diritti di **ognuno**
	PP	Pronome possessivo	*mio, tuo, suo, loro, proprio*	il **mio** qui pi bella della **loro**

CPoS	FPoS	Descrizione	Esempi	Contesti d'uso
	PQ	Pronome interrogativo	*che, chi, quanto*	non so **chi** parta **quanto** costa? **che** ha fatto ieri?
	PR	Pronome relativo	*che, cui, quale*	**ci** che dice il **quale** afferma a **cui** parlo
	PC	Pronome clitico	*ci, vi, mi, ti, la, le*	**lo** vidi **li** ho sentiti aver**la** **le** dissero, **le** videro **mi** dicono **ci** sposiamo **vi** credo **si** sente, **si** sentono **ci** vado spesso
R	RD	Articolo determinativo	*il, lo, la, i, gli, le*	**il** libro **i** gatti
	RI	Articolo indeterminativo	*uno, un, una*	**un** amico **una** bambina
S	S	Nome comune	*amico, insegnante, verità*	l'**amico** la **verità**
	SA	Abbreviazione	*ndr, a.C., d.o.c., km*	30 **km** sesto secolo **a.C.**
	SP	Nome proprio	*Monica, Pisa, Fiat, Sardegna*	**Monica** scrive
T	T	Predeterminante	*tutto, entrambi, ambedue*	**tutte** le notizie **ambedue** le idee
V	VA	Verbo ausiliare	*avere, essere, venire*	il peggio **è**passato **ho** scritto una lettera **viene** fatto domani
	VM	Verbo modale	*volere, potere, dovere, solere*	non **posso** venire **vuole** il libro
	V	Verbo	*mangio, avere, passato, camminando*	il peggio è **passato** ho **scritto** una lettera **vengo** domani
X	X	Residuo	include formule, parole sconosciute, simboli alfabetici e simili	distanziare di **43"** mi **piacce**

Tabella 8.1: Le categorie morfosintattiche.

Lo schema di annotazione sintattica a dipendenze

Lo schema di annotazione sintattica a dipendenze, riportato nella Tabella 8.2, comprende 29 relazioni di dipendenza che legano una testa sintattica (segnalata in grassetto nei periodi di esempio) al suo token dipendente (sottolineato e in grassetto negli esempi).

Etichetta	**Tipo di relazione**	**Descrizione**	**Esempi**
Arg	argument	Relazione tra una testa verbale o nominale e una frase completiva non soggetto (sia essa infinitiva o meno).	*Il 63% dei francesi ha* ***imposto*** *al presidente* ***di*** *rinunciare alla sua bomba* *È giunto il* ***momento*** ***di*** *creare un'area denuclearizzata* *Le autorità hanno* ***annunciato*** ***che*** *il blitz è concluso* *La* ***decisione*** ***di*** *continuare...* ***escludendo*** ***che*** *il militare volesse veramente mettere in pericolo...* *si sono* ***rifiutati*** ***di*** *fornire informazione*
Aux	auxiliary	Relazione tra una testa verbale e il suo ausiliare.	*Il corazziere* ***è*** ***stato*** *individuato* *Il corazziere è* ***stato*** ***individuato*** ***Ha*** ***dichiarato*** *di* ***aver*** ***pagato*** *i terroristi*
clit	clitic	Relazione tra un pronome clitico e una testa verbale usata in forma pronominale.	*La sedia* ***si*** *è* ***rotta*** *Non* ***ci*** ***rendiamo*** *conto* ***Si*** ***tratta*** *della scoperta*
comp	complement	Relazione tra una testa e un complemento preposizionale, sia esso modificatore o argomento. Questa relazione funzionale sottospecificata è particolarmente utile in quei casi in cui è difficile stabilire la natura argomentale o di modificatore del complemento.	*Fu* ***assassinata*** ***da*** *un pazzo* *E' più* ***interessante*** ***del*** *libro* ***Oggi*** ***come*** *allora* ***Più*** ***di*** *quattrocento esemplari* ***Osteggiata*** ***dal*** *governo di Berna* ***Grande*** ***quanto*** *mezza Italia*

Etichetta	**Tipo di relazione**	**Descrizione**	**Esempi**
comp_ind	indirect complement/object	Relazione sottotipo della relazione 'comp' circoscritta ai complementi di termine.	*Ho **dato** il libro **a** lui* *I carabinieri **gli** hanno **recapitato** il decreto*
comp_loc	locative complement	Relazione sottotipo della relazione 'comp' circoscritta ai complementi di luogo, sia esso di stato o di moto.	*Si **trovava** **in** un parco* *Era **uscito** **di** casa alle 10*
comp_temp	temporal complement	Relazione sottotipo della relazione 'comp' circoscritta ai complementi di tempo.	***Nel** 1985 è stata **uccisa** un'antropologa* *L'allarme è **scattato** la scorsa **settimana***
con	copulative conjunction	Relazione tra un elemento congiuntivo, sia esso una congiunzione coordinativa o altro, e il primo elemento coordinato all'interno di una struttura coordinativa (testa dell'intera struttura).	*Una ragazza **violentata** **e** sequestrata da due slavi* ***Gabriella** **e** Paolo sono partiti* *Hanno **riarmato,** addestrato e preparato l'esercito* *Hanno riarmato, **addestrato** **e** preparato l'esercito* ***Scontri,** assalti e centinaia di feriti* *Scontri, **assalti** **e** centinaia di feriti*
concat	concatenation	Relazione tra due tokens che costituiscono un'unità poliremática fissa tipicamente usata per nomi di società, nomi propri, ecc. . .).	*Il segretario di **De Michelis*** *L'enciclica **Mulieris dignitatem*** *La **International Public** Sport* *La International **Public Sport***

Etichetta	Tipo di relazione	Descrizione	Esempi
conj	conjunct linked by a copulative conjunction (con)	Relazione tra il secondo (o il terzo, quarto, ecc..) elemento parte di una struttura coordinata e il primo token, il quale rappresenta la testa sintattica dell'intera struttura. È usata sempre in coppia con la relazione 'con'.	*Una ragazza* ***violentata*** *e* ***sequestrata*** *da due slavi* ***Gabriella*** *e* ***Paolo*** *sono partiti* *Hanno* ***riarmato***, ***addestrato*** *e preparato l'esercito* *Hanno riarmato,* ***addestrato*** *e* ***preparato*** *l'esercito* ***Scontri***, ***assalti*** *e centinaia di feriti*
det	determiner	Relazione tra una testa nominale e il suo determinante.	***Una*** ***sala*** *ha dovuto essere sgomberata* *Rilevata* ***la*** ***presenza*** *di gas*
dis	disjunctive conjunction	Relazione tra un elemento disgiuntivo, sia esso una congiunzione disgiuntiva o altro, e il primo elemento coordinato all'interno di una struttura coordinativa (testa dell'intera struttura).	*Cassonetti dell'immondizia* ***rovesciati*** ***o*** *incendiati* *Partecipa* ***a*** *manifestazioni politiche* ***o*** *a dibattiti*
disj	conjunct in a disjunctive compound linked by a disjunctive conjunction (dis)	Relazione che unisce il secondo (o il terzo, quarto, ecc..) elemento parte di una struttura coordinata al primo token, il quale rappresenta la testa sintattica dell'intera struttura. È usata sempre in coppia con la relazione 'dis'.	*Cassonetti dell'immondizia* ***rovesciati*** *o* ***incendiati*** *Partecipa* ***a*** *manifestazioni politiche o* ***a*** *dibattiti*

Etichetta	**Tipo di relazione**	**Descrizione**	**Esempi**
mod	modifier	Relazione tra una testa e il suo modificatore; tale relazione copre modificatori di tipo frasale, aggettivale avverbiale e nominale.	*I colori* ***sono sempre*** *gli stessi* ***Colori intensi*** *Trionfo di Didoni nei* ***20 km*** *di marcia* ***Cesare*** *l'****Imperatore*** ***Per*** *arrivare in tempo, sono* ***partito*** *molto presto* ***Quando*** *la campanella suona, i bambini* ***escono*** *da scuola*
mod_loc	locative modifier	Relazione sottotipo della relazione 'mod' circoscritta ai modificatori con valore di moto o stato in luogo.	*Non* ***so dove*** *Tutto* ***cominciò*** *proprio* ***lì*** *Avrei voluto* ***fermar****mi* ***qui*** *più a lungo*
mod_rel	relative modifier	Relazione la testa verbale di una frase relativa e il suo antecedente. Lo stesso tipo di relazione è usato nel caso delle relative libere per collegare la testa verbale della relativa al pronome *chi*.	***Box*** *che è stato* ***trovato*** *nel pomeriggio* *Quell'****ordine*** *che i due Stranamore pentiti avevano* ***imposto*** *per cinquant'anni* *Non è mai stato accertato* ***chi volle*** *la sua morte*
mod_temp	temporal modifier	Relazione sottotipo della relazione 'mod' circoscritta ai modificatori con valore temporale.	***Ieri*** *hanno* ***dormito*** *all'aperto* *Scoperto 75* ***anni fa*** *Non* ***superano mai*** *gli 8 milioni*
modal	modal verb	Relazione tra una testa verbale e un verbo modale.	*Una sala ha* ***dovuto*** *essere* ***sgomberata*** *Avrebbe* ***potuto ripetersi***
neg	negative	Negazione (*no* o *non*).	*A volte* ***non dormo***
obj	direct object	Relazione tra un predicato e il suo oggetto diretto (sempre non–frasale).	***Hanno*** *un* ***modo*** *di ragionare rozzo?* ***Centellinando*** *le* ***informazioni*** *È giunto il momento di* ***creare*** *un'****area*** *denuclearizzata* ***Rilevata*** *la* ***presenza*** *di gas*

Etichetta	**Tipo di relazione**	**Descrizione**	**Esempi**
pred	predicative complement	Relazione tra una testa verbale e un complemento predicativo sia esso del soggetto o dell'oggetto.	*L'incontro è* ***stato fatale*** *Questo* ***è*** *il* ***messaggio*** *finale*
pred_loc	locative predicate	Relazione sottotipo della relazione 'pred' circoscritta ai complementi predicativi con valore di stato in luogo.	*Il presidente non* ***era in*** *casa*
pred_temp	temporal predicate	Relazione sottotipo della relazione 'pred' circoscritta ai complementi predicativi con valore temporale.	*La riunione* ***è alle*** *5*
prep	preposition	Relazione tra una testa preposizionale e il suo complemento, sia esso frasale o meno.	*Un contributo* ***alla lotta*** *contro la criminalità* *Un contributo alla lotta* ***contro*** *la* ***criminalità*** *Prima* ***di partire*** *ho telefonato*
punc	punctuation	Relazione tra un token-parola e un token-segno di punteggiatura.	***Teatro*** *della tragedia* **,** *...*
ROOT	sentence root	Radice del periodo.	***Desidero*** *dormire* *Note that only the dependent is shown, since the head is a fictitious root node*
sub	subordinate clause	Relazione tra una congiunzione subordinativa e la testa verbale di una frase subordinata.	*Ha detto* ***che*** *non* ***intendeva*** *fare nulla* *Le autorità hanno annunciato* ***che*** *il blitz è* ***concluso*** *Venne ucciso* ***mentre cercava*** *di difendere la ragazza*
subj	subject	Relazione tra una testa verbale di forma attiva e il suo soggetto, sia esso frasale o meno.	*il* ***testimone*** *ha* ***parlato*** *subito* *le* ***vittime seguivano*** *gli aiuti* *È* ***stato*** *facile* ***ricostruire*** *le telefonate in partenza dal portatile*

Etichetta	**Tipo di relazione**	**Descrizione**	**Esempi**
subj_pass	passive subject	Relazione tra una testa verbale di forma passiva e il suo soggetto.	*I **missionari** erano stati **rapiti** la mattina presto* *Circa 83.000 **franchi** furono **spesi***

Tabella 8.2: Le relazioni di dipendenza.

Appendice II

Nelle tabelle di questa seconda appendice è riportato l'elenco dei frames annotati durante il caso di studio descritto nel Capitolo 7 e qui organizzati nel modo seguente:

- nella Tabella 8.3, per ogni frame deontico è riportata la rispettiva LU evocatrice (prima colonna), seguita dal numero di istanze annotate (seconda colonna), e il frame che ne rappresenta il 'dovere', 'permesso' o 'divieto' normato (terza colonna) con la relativa LU evocatrice (quarta colonna);
- nelle Tabelle 8.4 e 8.5 sono riportate le annotazioni dei nuovi frames GRANT_LEGAL_PERMISSION e VIOLATION proposti;
- nella Tabella 8.6 sono riportate le annotazioni dei frames non deontici, ma considerati importanti per la piena descrizione dello scenario di 'obbligo'.

Per ogni LU poliremática è riportato tra parentesi tonde il verbo o la preposizione supporto. I casi in cui esse siano state acquisite in modo automatico sulla base della metodologia illustrata nel Paragrafo 7.3.1 sono segnalati in grassetto. Le parentesi quadre sono state utilizzate per segnalare i casi di riconoscimento automatico solo di una parte della LU poliremática. È il caso di *[entrata]**(in)vigore*** dove la fase di estrazione automatica ha individuato *in vigore* come unità poliremática; la revisione manuale ha poi messo in luce che, non solo la preposizione *in*, ma anche il sostantivo *entrata* funge da 'supporto' alla capacità evocatrice del termine *vigore*.

LU	No.	Frame regolato	LU
Frame deontico: Being_obligated			
tenuto	6	ADOPT_SELECTION	*adozione, adottare*
		ACTIVITY_START	*procedere*
		Unknown_2	*appaltare*
		TELLING	*(dare)comunicazione, aggiornare, informare*
		FINING	*pagamento*
obbligato	13	Unknown_11	*(costituire)garanzia, (costituire)garanzia fidejussoria*
		COMPLIANCE	*rispetto*
		CAUSE_TO_PERCEIVE	*esporre*
		RECEIVING	*accettare*
		ADOPT_SELECTION	*adottare*
		STORING	*mantenere, conservare*
		Unknown_7	*pagamento*
		SECRECY_STATUS	*segreto*
		GIVING	*conferire*
		PARTICIPATION	*partecipare*
		TELLING	*trasmettere*
(sottoposto all')obbligo	1	PUBLIC_SERVICES	*servizio pubblico*
(essere)soggetto	1	FINING	*sanzione amministrativa*
(essere soggetto all')obbligo	1	DOCUMENTS	*registro*
chiamato	1	CAUSE_TO_BE_INCLUDED	*integrare*
(avere)obbligo	8	COLLABORATION	*cooperare*
		GRANT_PERMISSION	*consentire*
		Unknown_4	*tenere indenne*
		FINING	*risarcire*
		TELLING	*sottoporre, trasmettere, presentare*
		SUPPLY	*fornire*
		LOCATING	*individuare*
		DECIDING	*determinare*
(assumere)obbligo	1	INSPECTING	*verificare*
obbligo	2	Unknown_7	*versamento*
		Unknown_6	*compensazione*
Frame deontico: Being_obligatory			
obbligatorio	6	Unknown_10	*caratterizzazione di base*
		DOCUMENTS	*procedimento di valutazione di impatto ambientale*
		Unknown_12	*sistema*
		Unknown_9	*misurazione*

LU	No.	Frame regolato	LU
		USING	*utilizzo*
Frame deontico: Imposing_obligation			
(fare)obbligo	6	COMPLIANCE	*osservare, adeguare, rispetto*
		GIVING	*affidare, conferire*
		RECEIVING	*accettare*
		SUPPLY	*dotare*
irrogato	2	REWARDS_AND_PUNISHMENT	*sanzione, sanzione*
disporre	1	ACTIVITY_STOP	*cessazione*
prevedere	1	INTENTIONALLY_CREATE	*istituzione*
		STORING	*raccolta*
		DESTROYING	*smaltimento*
		TRANSFER	*conferimento*
Frame deontico: Permitting			
autorizzato	5	HINDERING	*limite*
		Unknown_7	*spesa*
		CAUSE_CHANGE	*trasformazione*
		PARTICIPATION	*partecipazione*
permesso	2	Unknown_8	*immissione sul mercato*
		RECORDING	*immatricolazione*
concessione	3	CREATING	*coltivazione*
		CAUSE_FLUIDIC_MOTION	*derivazione, derivazione*
consentito	1	Unknown_8	*immissione sul mercato*
concesso	1	DOCUMENTS	*autorizzazione*
Frame deontico: Prohibiting			
interdizione	1	Unknown_5	*traffico veicolare*
(fare)divieto	3	COMMERCE_SCENARIO	*commercializzare*
		HINDERING	*(introdurre)restrizione*
		DISPERSAL	*disperdere*
divieto	5	COMMERCE_SCENARIO	*commercializzazione*
		CAUSE_FLUIDIC_MOTION	*scarico*
		INSTALLING	*istallazione*
		USING	*uso*
		PLACING	*introduzione*
vietato	12	CAUSE_CHANGE	*trasformazione, conversione*
		CUTTING	*taglio*
		Unknown_8	*immissione sul mercato*
		BUILDING	*realizzazione*
		ACTIVITY_START	*attività*
		AGRICULTURE	*pascolo, stabulazione*
		HUNTING	*caccia*
		USING	*utilizzo*
		DESTROYING	*coincenerimento, smaltimento*

LU	**No.**	**Frame regolato**	**LU**
		DISPERSAL	*diffusione*
Frame deontico: Deny_permission			
(fare)divieto	1	ACTIVITY_START	*procedere*
interdizione	1	LEADERSHIP	*ufficio direttivo*
negare	1	DOCUMENTS	*autorizzazione*
proibire	2	USING	*uso*
		COMMERCE_SELL	*vendita*

Tabella 8.3: Le istanze di frames deontici e regolati annotate.

LU	No.	Frame regolato	LU
Frame: Grant_legal_permission			
(dare)facoltà	1	IMPOSING_OBLIGATION	*disporre*
autorizzare	1	Unknown_8	*immissione sul mercato*
autorizzato	1	CAUSE_CHANGE	*(apportare)variazione*

Tabella 8.4: Le istanze del nuovo frame GRANT_LEGAL_PERMISSION proposto.

Frame	LU	No. istanze
VIOLATION	*(in)violazione*	3
	violazione	2

Tabella 8.5: Le istanze del nuovo frame VIOLATION proposto.

Frame	LU	No. istanze
COMPLIANCE	***(in)conformità***	2
	conformare	3
	ottemperare	1
BEING_IN_ EFFECT	*(avere)validità*	1
	(acquistare)efficacia	1
	(mantenere)efficacia	1
	(conservare)efficacia	1
	valere	1
	[entrata] ***(in)vigore***	2
	[entrare] ***(in)vigore***	1
	[rimanere] ***(in)vigore***	1
	[restare] ***(in)vigore***	3
ENFORCING	*(in)applicazione*	1
REQUIRED_EVENT	*dovere*	1
	(fare)obbligo	1

Frame	LU	No. istanze
Law	*provvedimento*	1

Tabella 8.6: Le istanze dei frames rilevanti per la descrizione dello scenario di 'obbligo'.

Bibliografia

T. Agnoloni, L. Bacci, F. Francesconi, W. Peters, S. Montemagni e G. Venturi. A two-level knowledge approach to support multilingual legislative drafting. In P. Casanovas J. Breuker e M.C.A. Klein, editori, *Law and the Semantic Web. Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications*, LNCS, vol. 188, pp. 177–198. Springer–Verlag, Berlin Heidelberg, 2009.

B.T.S. Atkins, C.J. Fillmore e C.R. Johnson. Lexicographic relevance: Selecting information from corpus evidence. In T. Fontanelle, editore, *FrameNet and Frame Semantics*, volume 16(3), pp. 251–280. International Journal of Lexicography, Special Issue, 2003a.

B.T.S. Atkins, M. Rundell e H. Sato. The contribution of framenet to practical lexicography. In T. Fontanelle, editore, *FrameNet and Frame Semantics*, volume 16(3), pp. 333–357. International Journal of Lexicography, Special Issue, 2003b.

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.

C.F. Baker e C. Fellbaum. Wordnet and framenet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09)*, pp. 125–129, Suntec, Singapore, 2009.

C.F. Baker, C.J. Fillmore e J.B. Lowe. The berkeley framenet project. In *Proceedings of the 36th ACL Meeting and 17th ICCL Conference*. Morgan Kaufmann, 1998.

R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli e C. Soria. Automatic classification and analysis of provisions in legal texts: a case study. In

A. R. Meersman, Z. Tari e A. Corsaro, editori, *On the Move to Meaningful Internet Systems (OTM 2004 Workshops)*, LNCS, vol. 3292, pp. 593–604. Springer–Verlag, 2004.

R. Basili, A. Moschitti, M-T. Pazienza e F.M. Zanzotto. A contrastive approach to term extraction. In *Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA 2001)*, Nancy, France, 2001.

G.L. Beccaria. Linguaggi settoriali e lingua comune. In G.L. Beccaria, editore, *I linguaggi settoriali in Italia*, pp. 7–59. Milano, Bompiani, 1973.

P. Bellucci. *A onor del vero. Fondamenti di linguistica giudiziaria italiana.* Torino, UTET, 2005.

A. Belvedere. Linguaggio giuridico. In *Digesto delle discipline privatistiche, Sezione civile*, volume XI, pp. 21–31, 1994a.

A. Belvedere. Il linguaggio del codice civile: alcuni osservazioni. In U. Scarpelli e P. Di Lucia, editori, *Il linguaggio del diritto*, pp. 403–452. Milano, LED, 1994b.

A. Belvedere. I poteri semiotici del legislatore (alice e l'art. 12 preleggi). In L. Gianformaggio e alii, editori, *Scritti per Uberto Scarpelli*, pp. 85–103. Milano, Giuffrè, 1998.

A. Belvedere. Pragmatica e semantica nell'art.12 preleggi. In D. Veronesi, editore, *Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati. Atti del Convegno di studi (Bolzano, 1–3 ottobre 1998)*, pp. 49–58. Unipress, Padova, 2000.

L. Bentivogli, A. Bocco e E. Pianta. Archiwordnet: Integrating wordnet with domain–specific knowledge. In *Proceedings of the Second Global WordNet Conference*, pp. 39–46, Brno, Czech Republic, 2004.

F. Bertagna, M-T. Sagri e D. Tiscornia. Jur–wordnet. In *Proceedings of the Second International WordNet Conference (GWC 2004)*, pp. 305–310, Brno, Czech Republic, 2004.

C. Biagioli. Legimatica: verso una seconda generazione. In C. Biagioli, P. Mercatali e G. Sartor, editori, *Legimatica. Informatica per legiferare*, pp. 75–91. Napoli, ESI, 1995.

C. Biagioli. *Modelli funzionali delle leggi. Verso testi legislativi autoesplicativi*, volume 6 of *Legal Information and Communication technologies*. European Press Academic Publishing, 2009.

C. Biagioli e S. Pietropaoli. Considerazioni sulle tecniche di costruzione delle disposizioni normative nella prassi legislativa italiana. sanzioni e obblighi esplicitamente sanzionati: un caso affrontato nello studio per l'evidenziazione automatica della metainformazione nir, finalizzata all'annotazione elettronica dei testi in rete. In *Informatica e diritto*, volume XXIX(1–2), pp. 77–98, 2003.

C. Biagioli, G. Bianucci, P. Mercatali e D. Tiscornia. Introduzione. l'analisi automatica dei testi giuridici e politici. In P. Mercatali, editore, *Computer e linguaggi settoriali. Analisi automatica di testi giuridici e politici*, pp. 15–27. Milano, Franco Angeli, 1988a.

C. Biagioli, P. Mercatali e D. Tiscornia. Le formule per l'analisi automatica della leggibilità: la formula di flesch per il controllo di documenti giuridici. In P. Mercatali, editore, *Computer e linguaggi settoriali. Analisi automatica di testi giuridici e politici*, pp. 45–99. Milano, Franco Angeli, 1988b.

D. Biber. Using register–diversified corpora for general language studies. In *Computational Linguistics Journal*, volume 19(2), pp. 219–241, 1993.

D. Biber, S. Conrad e R. Reppen. *Corpus linguistics. Investigating Language Structure and Use*. Cambridge University Press, 1998.

H.C. Boas. Bilingual framenet dictionaries for machine translation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pp. 1364–1371, Las Palmas, Spain, 2002. European Language Resources Association (ELRA).

H.C. Boas, editore. *Multilingual FrameNets in computational lexicography: methods and applications*. Mouton de Guyter, 2009.

N. Bobbio. Scienza del diritto e analisi del linguaggio. In U. Scarpelli, editore, *Diritto e analisi del linguaggio*, pp. 287–324. Milano, Edizioni di Comunità, 1976.

O. Bodenreider. Lexical, terminological and ontological resources for biological text mining. In S.Ananiadou e J. McNaught, editori, *Text mining for biology and biomedicine*, pp. 43–66. Boston, Artech House, 2006.

A. Bolioli, L. Dini, P. Mercatali e F. Romano. For the automated mark–up of italian legislative texts in xml. In *Proceedings of Legal Knowledge and Information Systems (JURIX) Conference*, pp. 21–30, London, United Kingdom, 2002. IOS Press.

F. Bonin, F. Dell'Orletta, S. Montemagni e G. Venturi. Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio. In *Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010)*, pp. 207–220, 27-29 settembre, Viterbo, 2010a.

F. Bonin, F. Dell'Orletta, S. Montemagni e G. Venturi. A contrastive approach to multi–word extraction from domain–specific corpora. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 3222–3229, La Valletta, Malta, 2010b. European Language Resources Association (ELRA).

S. Bonzi. Syntactic patterns in scientific sublanguages: a study of four disciplines. In *Journal of the American Society for Information Science*, volume 41(2), pp. 121–131, 1990.

L. Borin, D. Dannélls, M. Forsberg, M. Toporowska Gronostaj e D. Kokkinakis. Thinking green: Toward swedish framenet++. In *Proceedings of the FrameNet Masterclass and Workshop*, Università Cattolica, Milano, 2009.

G. Bouma, G. van Noord e R. Malouf. Alpino: Wide–coverage computational analysis of dutch. In W. Daelemans, K. Sima'an, J. Veenstra e J. Zavrel, editori, *Computational Linguistics in the Netherlands*, pp. 45–59. CLIN Meeting, Rodopi, Amsterdam, 2000.

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.

J. Breuker. Dreams and awakenings about legal ontologies. In *Intervento tenuto in occasione del 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 09)*, Barcelona, Spain, 2009.

J. Breuker e R. Hoekstra. Epistemology and ontology in core ontologies: Folaw and lri-core, two core ontologies for law. In *Proceedings of the Workshop on Core Ontologies in Ontology Engineering (EKAW04)*, pp. 15–27, Northamptonshire, UK, 2004.

J. Breuker, R. Hoekstra, A. Boer, K. van den Berg, R. Rubino, G. Sartor, M. Palmirani, A. Wyner e T. Bench-Capon. Owl ontology of basic legal concepts (lkif-core). In *Deliverable 1.4 D.1.4, ESTRELLA project (IST-2004-027655)*, 2007.

P. Buitelaar e B. Sacaleanu. Extending synsets with medical terms. In *Proceedings of the First International WordNet Conference*, Mysore, India, 2002.

P. Buitelaar, P. Cimiano e B. Magnini. Ontology learning from text: an overview. In Buitelaar et al., editore, *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications, vol. 123, pp. 3–12. Springer–Verlag, Berlin Heidelberg, 2005.

P. Buitelaar, P. Cimiano, P. Haase e M. Sintek. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web Research (ESWC 2009)*, pp. 111–125, Heraklion, Crete, Greece, 2009.

A. Burchardt, A. Frank e M. Pinkal. Building text meaning representations from contextually related frames. a case study. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS–6)*, Tilburg, The Netherlands, 2005.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado e M. Pinkal. Framenet for the semantic analysis of german: Annotation, representation and automation. In H.C. Boas, editore, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pp. 209–244. Mouton de Guyter, 2009.

N. Casellas. *Legal Ontology Engineering. Methodologies, Modelling Trends, and the Ontology of Professional Judicial Knowledge.* Law, Governance and Technology Series (Vol. 3). Springer–Verlag, Berlin/Heidelberg, 2011.

S. Cassese. Introduzione allo studio della normazione. In *Rivista trimestrale di diritto pubblico*, volume 2, pp. 307–330, 1992.

V.R. Charrow, J.A. Crandall e R.P. Charrow. Characteristics and functions of legal language. In R. Kittredge e J. Lehrberger, editori, *Sublanguage: Studies of Language in Restricted Semantic Domains)*, pp. 177–190. deGruyter, Berlin, 1982.

W-C. Chou, R.T-H. Tsai, Y-S. Su, W. Ku, T-Y. Sung e W-L. Hsu. A semi-automatic method for annotating a biomedical proposition bank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pp. 5–12, Sydney, Australia, 2009.

T.M. Chung e P. Nation. Identifying technical vocabulary. volume 32) of *System*, pp. 251–263, 2004.

K.W. Church e P. Hanks. Word association norms, mutual information, and lexicography. volume 16(1) of *Computational Linguistics*, pp. 22–29, 1990.

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.

K.B. Cohen, M. Palmer e L. Hunter. Nominalization and alternations in biomedical language. In *PLoS ONE*, volume 3(9), pp. 1–21, 2008.

M. Collins. *Head–Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, 1999.

M-E. Conte. Epistemico, deontico, anankastico. In A. Giacalone e G. Crocco Galéas, editori, *From Pragmatics to Syntax. Modality in Second Language Acquisition*, pp. 3–9. Narr, Tubingen, 1995.

M. Cortelazzo. Lingua e diritto in italia. il punto di vista dei linguisti. In L. Schena, editore, *La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno Internazionale, Milano, 5–6 ottobre*, pp. 35–50. Roma, Cisu (Centro d'Informazione e Stampa Universitaria), 1997.

E. de Maat e R. Winkels. Formal models of sentences in dutch law. In *Proceedings of the Workshop Applying Human Language Technology to the Law*, pp. 28–40, Pittsburgh, Pennsylvania, 2011.

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.

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 Società di Linguistica Italiana (SLI 2010)*, 27-29 settembre, Viterbo, 2010a.

F. Dell'Orletta, A. Lenci, S. Marchi, S. Montemagni e V. Pirrelli. Text-2-knowledge: una piattaforma linguistico–computazionale per l'estrazione di conoscenza da testi. In *Atti del XL Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2006)*, 20–28 settembre, Vercelli, 2006.

F. Dell'Orletta, S. Montemagni, E.M. Vecchi e G. Venturi. Tecnologie linguistico–computazionali per il monitoraggio delle competenze linguistiche di apprendenti l'italiano come l2. In *Comunicazione tenuta al convegno IT.L2: italiano lingua seconda nell'università, nella scuola e sul territorio*, 12–13 novembre, Vercelli, 2010b.

F. Dell'Orletta, S. Montemagni e G. Venturi. Readit: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pp. 73–83, Edinburgh, Scotland, 2011.

T. DeMauro. *Storia linguistica dell'Italia unita.* Bari, Laterza, 1963.

T. DeMauro. Introduzione. il linguaggio della *Costituzione*. In *Costituzione della Repubblica Italiana (1947)*, pp. vii–xxxii. Torino, UTET, 2006.

T. DeMauro e M. Voghera. Scala mobile. un punto di vista sui lessemi complessi. In P. Benincà et al., editore, *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, pp. 99–131. Roma, Bulzoni, 1996.

J. Dinarelli, E. Pianta, S. Vrochidis e S. Papadopoulos. Integration of semantic, metadata and image search engines with a text search engine for patent retrieval. In *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008)*, Tenerife, Spain, 2008.

M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti e G. Riccardi. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of the EACL Workshop on Semantic Representation of Spoken Language*, Athens, Greece, 2009.

A. Dolbey. *BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology*. Ph.D. thesis, University of California, Berkeley, 2009.

T. Dunning. Accurate methods for the statistics of surprise and coincidence. volume 19(1) of *Computational Linguistics*, pp. 61–74, 1993.

P. Eklund-Braconi, editore. *Il linguaggio normativo delle Comunità Europee. Studi quantitativi e semantici sul lessico con particolare riguardo al concetto di ambiente*. Dipartimento di francese e italiano - Università di Stoccolma, Stockolm, Graphium, 2000.

E. Ellsworth, K. Erk, P. Kingsbury e S. Pado. Propbank, salsa and framenet: How design determines product. In *Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora in conjunction with LREC 2004*, pp. 17–23, Lisbon, Portugal, 2004.

K. Erk e S. Padó. Shalmaneser – a flexible toolbox for semantic role assignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006.

K. Erk, A. Kowalski e S. Pado. The salsa annotation tool–demo description. In *Proceedings of the 6th Lorraine-Saarland Workshop*, pp. 111–113, Nancy, France, 2003.

C. Fellbaum. English verbs as a semantic net. In *International Journal of Lexicography*, volume 3(4), pp. 40–61, 1993b.

C. Fellbaum, editore. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

C. Fellbaum, D. Gross e K. Miller. Adjectives in wordnet. In *International Journal of Lexicography*, volume 3(4), pp. 26–39, 1993a.

C.J. Fillmore. Scenes–and–frames semantics. In A. Zampolli, editore, *Linguistic Structures Processing*, pp. 55–81. Dordrecht: North Holland Publishing, 1977.

C.J. Fillmore. Frame semantics. In Linguistic Society of Korea, editore, *Linguistics in the Morning Calm*, pp. 111–138, Seoul, Hanshin, 1982.

C.J. Fillmore. Frame and the semantics of understanding. In *Quaderni di semantica*, volume IV(2), dicembre, pp. 222–254, 1985.

C.J. Fillmore e B.T. Atkins. Toward a frame–based lexicon: the semantics of risk and its neighbors. In A. Lehrer e E.F. Kittay, editori, *Frames, Fields and Contrasts*, pp. 75–102. Lawrence Erlbaum Associates Publishers, Hillsdale, 1992.

C.J. Fillmore e B.T.S. Atkins. Starting where the dictionaries stop: The challenge for computational lexicography. In B.T.S. Atkins e A. Zampolli, editori, *Computational Approaches to the Lexicon*, pp. 349–393. Oxford, Oxford University Press, 1994.

C.J. Fillmore e C.F. Baker. Frame semantics for text understanding. In *Proceedings of the WordNet and Other Lexical Resources Workshop, in conjunction with NAACL*, Pittsburgh, Pennsylvania, 2001.

C.J. Fillmore e C.F. Baker. A frames approach to semantic analysis. In B. Heine e H. Narrog, editori, *The Oxford Handbook of Linguistic Analysis*, pp. 313–339. Oxford University Press, 2010.

C.J. Fillmore, C.F. Baker e H. Sato. Framenet as a 'net'. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1091–1094, Lisbon, Portugal, 2004.

C.J. Fillmore, S. Narayanan e C. Baker. What can linguistics contribute to event extraction? In *Proceedings of the Twenty-First National Conference on Artificial Intelligence Workshop on Event Extraction and Synthesis (AAAI'06)*, Boston, Massachusetts, 2006.

P. Fiorelli. Premessa. In P. Mariani Biagini, editore, *Indice della lingua legislativa italiana. Inventario lessicale dei cento maggiori testi di legge tra il 1723 e il 1973*, pp. V–XII. Istituto per la Documentazione Giuridica del Consiglio Nazionale delle Ricerche, 1993.

P. Fiorelli. *Intorno alle parole del diritto*. Milano, Giuffrè, 2008.

G. Fiorentino. Web usability e semplificazione linguistica. In F. Venier, editore, *Rete Pubblica. Il dialogo tra Pubblica Amministrazione e cittadino: linguaggi e architettura dell'informazione*, pp. 11–38, Perugia, Edizione Guerra, 2007.

E. Francesconi, S. Montemagni, W. Peters e D. Tiscornia, editori. *Semantic Processing of Legal Texts*. LNAI 6036. Springer–Verlag, Berlin Heidelberg, 2010.

K. Frantzi e S. Ananiadou. The c–value / nc value domain independent method for multi–word term extraction. volume 6(3) of *Journal of Natural Language Processing*, pp. 145–179, 1999.

P. Fung e C. Benfeng. Biframenet: Bilingual frame semantics resource construction by cross–lingual induction. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pp. 931–937, Geneva, Switzerland, 2004. Association for Computational Linguistics.

A. Gangemi, R. Navigli e P. Velardi. The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In *Proceedings of the International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003)*, pp. 820–838, Catania, Italia, 2003a.

A. Gangemi, M-T. Sagri e D. Tiscornia. Metadata for content description in legal information. In *Proceedings of the LegOnt Workshop on Legal Ontologies*, 2003b.

A. Gangemi, M-T. Sagri e D. Tiscornia. A constructive framework for legal ontologies. In V.R. Benjamins et al., editore, *Law and the Semantic Web*, pp. 97–124. Berlin Heidelberg, Springer–Verlag, 2005.

B. Mortara Garavelli. *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino, Einaudi, 2001.

B. Mortara Garavelli. Strutture testuali e stereotipi nel linguaggio forense. In P. Mariani Biagini, editore, *La lingua, la legge, la professione forense. Atti del convegno Accademia della Crusca (Firenze, 31 gennaio–1 febbraio 2002*, pp. 3–19. Milano, Giuffrè, 2003.

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

D. Gildea e D. Jurafsky. Automatic labeling of semantic roles. In *Computational Linguistics Journal*, volume 28(3), pp. 245–288. MIT Press, Cambridge, MA, 2002.

A-N. Giuglea e A. Moschitti. Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the 21st International Conference on Computational Linguistics (ACL 2006)*, pp. 929–936, Sydney, Australia, 2006.

R. Grishman e R. Kittredge, editori. *Analyzing language in restricted domains: sublanguage description and processing*. Hillsdale, NJ, Lawrence Erlbaum, 1986.

R. Grishman, N. Thanh Nhan, E. Marsh e L. Hirschman. Automated determination of sublanguage syntactic usage. In *Proceedings of the 10th International Conference on Computational Linguistics*, pp. 96–100, Stanford, California, 1984.

B. Hachey e C. Grover. Extractive summarisation of legal texts. In *Artificial Intelligence and Law*, volume 14(4), pp. 305–345. MIT Press, Cambridge, MA, 2006.

P. Hanks. Do word meanings exist? In *Computers and the Humanities*, volume 34, pp. 205–215. Kluwer Academic Publishers, 2000.

P. Hanks. Mapping meaning onto use. In M.H. Corréard, editore, *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S.Atkins*, pp. 156–198. EURALEX 2002, 2002.

P. Hanks. Mapping meaning onto use: a pattern dictionary of english verbs. In *Proceedings of the American Association for Corpus Linguistics Conference (AACL 2008)*, Provo, Utah, 2008.

P. Hanks e J. Pustejovsky. A pattern dictionary for natural language processing. In *Revue Francaise de linguistique appliquée*, volume 10(2), 2005.

S. Harabagiu e C.A. Bejan. A knowledge extraction framework for biomedical pathways. In *Proceedings of the AMIA Summits Translational Science*, pp. 1–5, 2010.

Z.S. Harris. *Mathematical Structures of Language*. New York, Wiley, 1968.

G. Hirst. Ontology and the lexicon. In *Handbook on Ontologies in Information Systems*, pp. 209–230. Springer, 2003.

R. Jackendoff. Twistin' the night away. volume 73 of *Language*, pp. 534–559, 1997.

M. Jori. Definizioni e livelli di discorso giuridico. In U. Scarpelli e P. Di Lucia, editori, *Il linguaggio del diritto*, pp. 367–386. Milano, LED, 1994.

M. Jori e A. Pintore, editori. *Manuale di teoria generale del diritto*. Torino, Giappichelli, 1995.

A. Kilgarriff. I don't believe in word senses. In *Computers and the Humanities*, volume 31(2), pp. 91–113. Kluwer Academic Publishers, 1997.

K. Kipper-Schuler. *VerbNet: A broad–coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, 2005.

R. Kittredge. Variation and homogeneity of sublanguages. In R. Kittredge e J. Lehrberger, editori, *Sublanguage: Studies of Language in Restricted Semantic Domains)*, pp. 107–137. deGruyter, Berlin, 1982.

D. Kokkinakis e G.M. Toporowska. Linking swefn++ with medical resources, towards a medframenet for swedish. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pp. 68–71, Los Angeles, California, 2010.

F. Kuhn. A description language for content zones of german court decisions. In *Proceedings of the Language Resources and Evaluation Conference (LREC2010), Workshop on Semantic Processing of Legal Texts (SPLeT 2010)*, pp. 1–7, La Valletta, Malta, 2010.

G. Lame. Using nlp techniques to identify legal ontology components: concepts and relations. In J. Breuker R. Benjamins, P. Casanovas e A. Gangemi, editori, *Law and the Semantic Web. Legal Ontologies, Methodolo-*

gies, Legal Information Retrieval, and Applications, LNCS, vol. 3369, pp. 169–184. Springer–Verlag, Berlin Heidelberg, 2005.

M. Lease e E. Charniak. Parsing biomedical literature. Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP'05), pp. 58–69, 2005.

J. Lehrberger. Sublanguage analysis. In R. Grishman e R. Kittredge, editori, *Analyzing language in restricted domains: sublanguage description and processing*, pp. 19–38. Hillsdale, NJ, Lawrence Erlbaum, 1986.

A. Lenci, S. Montemagni V. Pirrelli e G. Venturi. Ontology learning from italian legal texts. In P. Casanovas J. Breuker e M.C.A. Klein, editori, *Law and the Semantic Web. Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications*, LNCS, vol. 188, pp. 75–94. Springer–Verlag, Berlin Heidelberg, 2009.

L. Lesmo. The turin university parser at evalita 2009. In *Proceedings of Evaluation of NLP and Speech Tools for Italian (Evalita 2009)*, Reggio Emilia, Italy, 2009.

B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, 1993.

B. Levin e M. Rappaport Hovav. Lexical semantics and syntactic structure. In S. Lappin, editore, *The Handbook of Contemporary Semantic Theory*, pp. 487–507. Blackwell, Oxford, 1996.

J.B. Lowe, C.F. Baker e C.J. Fillmore. A frame–semantic approach to semantic annotation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, in conjunction with ANLP–97*, Washington, D.C., USA, 1997.

P.L.M. Lucatuorto. Intelligenza artificiale e diritto: Le applicazioni giuridiche dei sistemi esperti. In *Ciberspazio e Diritto*, volume 7(2), pp. 219–242, 2006.

P. Lucisano e M.E. Piemontese. *Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana*, volume 3 of *Scuola e Città*. 1988.

C. Macleod, R. Grishman, A. Meyers, L. Barrett e R. Reeves. Nomlex: A lexicon of nominalizations. In *Proceedings of the 8th International Congress of the European Association for Lexicography (EURALEX 1998)*, pp. 187–193, Liége, Belgium, 1998.

B. Magnini e G. Cavaglià. Integrating subject field codes into wordnet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 1413–1418, Athens, Greece, 2000.

B. Magnini, C. Strapparava, G. Pezzulo e A. Gliozzo. The role of domain information in word sense disambiguation. In *Natural Language Engineering, special issue on Word Sense Disambiguation*, volume 8(4), pp. 359–373. Cambridge University Press, 2002.

M.P. Marcus, M.A. Marcinkiewicz e B. Santorini. Building a large annotated corpus of english: the penn treebank. volume 19(2), pp. 313–330. MIT Press, 1993.

R. Marinelli, L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari e A. Zampolli. The italian parole corpus: an overview. In A. Zampolli et al., editore, *Computational Linguistics in Pisa*, XVI–XVII(1), pp. 401–421. Pisa–Roma, IEPI, 2003.

R. Marinelli, A. Roventini e A. Enea. Building a maritime domain lexicon: a few considerations on the database structure and the semantic coding. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 465–468, Lisbon, Portugal, 2004.

K.T. Maxwell, J. Oberlander e V. Lavrenko. Evaluation of semantic events for legal case retrieval. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2009)*, pp. 39–41, Barcelona, Spain, 2009.

A. Mazzei, D.P. Radicioni e R. Brighi. Nlp–based extraction of modificatory provisions semantics. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pp. 50–57, Barcelona, Spain, 2009.

L.T. McCarty. Deep semantic interpretations of legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law (ICAIL2007)*, Stanford, California, 2007.

L.T. McCarty. Remarks on legal text processing – parsing, semantics and information extraction. In *Proceedings of the Worskhop on Natural Language Engineering of Legal Argumentation (NaLEA2009)*, Barcelona, Spain, 2009.

T.L. McCarty. Reflections on taxman: An experiment in artificial intelligence and legal reasoning. In *Harvard Law Review*, volume 90, pp. 837–893, 1977.

D. McClosky e E. Charniak. Self–training for biomedical parsing. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, pp. 101–104, Columbus, Ohio, 2008.

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.

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.

P. Mercatali. Legimatica e redazione delle leggi. In C. Biagioli, P. Mercatali e G. Sartor, editori, *Legimatica. Informatica per legiferare*, pp. 37–74. Napoli, ESI, 1995.

P. Mercatali. Dodici anni di legimatica. da una parola a una disciplina. In *Iter Legis*, volume 6, pp. 97–114, 2004.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young e R. Grishman. Annotating noun argument structure for nombank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 803–806, Lisbon, Portugal, 2004.

G.A. Miller. Nouns in wordnet: A lexical inheritance system. In *International Journal of Lexicography*, volume 3(4), pp. 10–25, 1993a.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross e K. Miller. Introduction to wordnet: An on-line lexical database. In *International Journal of Lexicography*, volume 3(4), pp. 1–9, 1993b.

M. Minsky. A framework for representing knowledge. In P. Winston, editore, *The Psychology of Computer Vision*, pp. 211–277. New York, McGraw–Hill, 1975.

S. Montemagni. Tecnologie linguistico–computazionali per il monitoraggio della lingua italiana. Presentazione tenuta nell'ambito della Giornata di Studio "Lo stato della lingua. Il CNR e l'italiano nel terzo millennio", 2010.

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. Abeillé, editore, *Building and using Parsed Corpora*, Language and Speech Series, pp. 189–210. Kluwer, Dordrecht, 2003.

E. Mustafaraj, M. Hoof e B. Freisleben. Larc: Learning to assign knowledge roles to textual cases. In *Proceedings of the 19th Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pp. 370–375, Melbourne Beach, Florida, 2006. AAAI Press.

H. Nakagawa e T. Mori. Automatic term recognition based on statistics of compound nouns and their components. volume 9(2) of *Terminology*, pp. 201–219, 2003.

M. Nakamura, S. Nobuoka e A. Shimazu. Towards translation of legal sentences into logical forms. In K. Satoh et al., editore, *New Frontiers in Artificial Intelligence*, LNCS, vol. 4914, pp. 349–362. Springer–Verlag, Berlin Heidelberg, 2008.

I. Niles e A. Pease. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE 03)*, pp. 23–26, Las Vegas, 2003.

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.

J. Nivre. *Inductive Dependency Parsing*. Springer, 2006.

J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel e D. Yuret. The conll 2007 shared task on dependency parsing. In *Proceedings of the the EMNLP-CoNLL*, pp. 915–932, 2007a.

J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel e D. Yuret. The conll 2007 shared task on dependency parsing. Proceedings of the EMNLP-CoNLL 2007, pp. 915–932, 2007b.

N.F. Noy, M. Musen, J.L.V. Mejino e C. Rosse. Pushing the envelope: Challenges in a frame-based representation of human anatomy. In *Stanford Medical Informatics*, Technical Report, 2002.

J. Nystedt. Ricchezza (o povertà?) lessicale nei documenti italiani della cee. In G. Alfieri e A. Cassola, editori, *La "Lingua d'Italia". Usi pubblici e istituzionali, Atti del XXlX Congresso Internazionale di Studi della SLI (Malta, 3–5 novembre 1998)*, pp. 471–491. Roma, Bulzoni, 1999.

J. Nystedt. L'italiano nei documenti della cee: le sequenze di parole. In D. Veronesi, editore, *Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati. Atti del Convegno di studi (Bolzano, 1–3 ottobre 1998)*, pp. 273–284. Unipress, Padova, 2000.

K.H. Ohara, F. Seiko, O. Toshio, S. Ryoko e S. Hiroaki andI. Shun. The japanese framenet project: An introduction. In *Proceedings of the Fourth international conference on Language Resources and Evaluation (LREC'04). Workshop "Building Lexical Resources from Semantically Annotated Corpora"*, Lisbon, Portugal, 2004. European Language Resources Association (ELRA).

K. Pala, P. Rychlý e P. Šmerk. Morphological analysis of law texts. In *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2007)*, pp. 21–26, Brno, Masaryk University, 2007.

K. Pala, P. Rychlý e P. Šmerk. Automatic identification of legal terms in czech legal texts. In W. Peters E. Francesconi, S. Montemagni e D. Tiscornia, editori, *Semantic Processing of Legal Texts*, LNCS, vol. 6036, pp. 83–94. Springer–Verlag, Berlin Heidelberg, 2010.

R. Mochales Palau e M.F. Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the*

12th International Conference on Artificial Intelligence and Law (ICAIL 2009), pp. 98–107, Barcelona, Spain, 2009.

M. Palmer, D. Gildea e P. Kingsbury. The proposition bank: A corpus annotated with semantic roles. In *Computational Linguistics Journal*, volume 31(1), 2005.

M-T. Pazienza, A. Stellato e A. Tudorache. A bottom–up comparative study of eurowordnet and wordnet 3.0 lexical and semantic relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2293–2299, Marrakech, Morocco, 2008.

A. Penas, F. Verdejo e J. Gonzalo. Corpus–based terminology extraction applied to information access. Proceedings of the Corpus Linguistics 2001, pp. 458–465, 2001.

M.R.L. Petruck. Typological considerations in constructing a hebrew framenet. In H.C. Boas, editore, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pp. 183–208. Mouton de Guyter, 2009.

E. Pianta, L. Bentivogli e C. Girardi. Multiwordnet: Developing and aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pp. 293–302, Mysore, India, 2002.

M.E. Piemontese. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnodid, 1996.

M.E. Piemontese. Il linguaggio della pubblica amministrazione nell'italia d'oggi. aspetti problematici della semplificazione linguistica. In G. Alfieri e A. Cassola, editori, *La "Lingua d'Italia". Usi pubblici e istituzionali, Atti del XXlX Congresso Internazionale di Studi della SLI (Malta, 3–5 novembre 1998)*, pp. 269–292. Roma, Bulzoni, 1999.

M.E. Piemontese. Leggibilità e comprensibilità delle leggi italiane. alcune osservazioni quantitative e qualitative. In D. Veronesi, editore, *Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati. Atti del Convegno di studi (Bolzano, 1–3 ottobre 1998)*, pp. 103–117. Unipress, Padova, 2000.

M.E. Piemontese. Leggibilità e comprensibilità dei testi delle pubbliche amministrazioni: problemi risolti e problemi da risolvere. In S. Covino, editore, *La scrittura professionale. Ricerca, prassi, insegnamento., Atti del l Convegno di studi (Perugia, Università per stranieri, 23–25 ottobre 2000)*, pp. 119–130. Firenze, Olschki, 2001.

M.E. Piemontese e M.T. Tiraboschi. Leggibilità e comprensibilità dei testi della pubblica amministrazione. strumenti e metodologie di ricerca al servizio del diritto a capire testi di rilievo pubblico. In E. Zuanelli, editore, *Il diritto all'informazione in Italia*, pp. 225–246. Roma, Presidenza del Consiglio dei Ministri. Dipartimento per l'informazione e l'editoria, 1990.

B. Plank e G. van Noord. Grammar-driven versus data–driven: which parsing system is more affected by domain shifts? In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING 2010)*, pp. 25–33, Uppsala, Sweden, 2010.

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.

M. Poprat, E. Beisswanger e U. Hahn. Building a biowordnet by using wordnet's data formats and wordnet's software infrastructure: a failure story. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP '08)*, pp. 31–39, Columbus, Ohio, 2008.

M. Rathert. Comprehensibility in forensic linguistics – new perspectives for frame semantics. In P. Brandt e E. Fuß, editori, *Form, Structure, and Grammar. A Festschrift presented to Günther Grewendorf on occasion of his 60th birthday*, pp. 337–352. Berlin, Akademie, 2006.

A. Reimerink, M. Garcìa de Quesadaa e S. Montero-Martìneza. Contextual information in terminological knowledge bases: A multimodal approach. In *Journal of Pragmatics*, volume 42(7), pp. 1928–1950, 2010.

E.L. Rissland. Ai and legal reasoning. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI85)*, 1985.

F. Romano. Strumenti per l'analisi semantica di testi legislativi. In *sito web "LIUC Papers"*, volume 183(supplemento a dicembre 2005), pp. 44–51, 2005.

A. Roventini, A. Alonge, N. Calzolari, B. Magnini e F. Bertagna. Italwordnet: a large semantic database for italian. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 783–790, Athens, Greece, 2000.

G. Rovere. Sottocodici e registri in testi tecnici. occorrenze e cooccorenza. In *Rivista Italiana di Dialettologia*, volume XIII, pp. 135–160, 1989.

G. Rovere. *Capitoli di linguistica giuridica. Ricerche su corpora elettronici.* Alessandria, Edizioni dell'Orso, 2005.

J. Ruppenhofer, C. Sporleder, R. Morante, C.F. Baker e M. Palmer. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pp. 106–111, Boulder, Colorado, 2009.

J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson e J. Scheffczyk. *FrameNet II: Extended Theory and Practice.* 2010.

F. Sabatini. Introduzione. In P. Mariani Biagini, editore, *La lingua, la legge, la professione forense. Atti del convegno Accademia della Crusca (Firenze, 31 gennaio–1 febbraio 2002*, pp. XXIII–XXV. Milano, Giuffrè, 2003.

N. Sager, C. Friedman e M. Lyman, editori. *Medial language processing.* Addison–Wesley Publishing Company, 1987.

M-T. Sagri. Strumenti semantici per laccesso allinformazione giuridica: Giur–wordnet. volume XXVIII(2) of *Informatica e diritto*, pp. 185–210, 2002.

G. Sartor. Fundamental legal concepts: A formal and teleological characterization. In *Artificial Intelligence and Law*, volume 14, p. 101142. Springer–Verlag, Netherlands, 2006.

U. Scarpelli. *Contributo alla semantica del linguaggio normativo.* Torino, Memoria dell'Accademia delle Scienze, 1959.

U. Scarpelli. Semantica giuridica. In A. Azara e E. Eula, editori, *Digesto Italiano*, volume XVII, 1969.

U. Scarpelli, editore. *Diritto e analisi del linguaggio*. Milano, Edizioni di Comunità, 1976.

U. Scarpelli. La definizione nel diritto. In U. Scarpelli, editore, *Diritto e analisi del linguaggio*, pp. 183–197. Milano, Edizioni di Comunità, 1976b.

J. Scheffczyk, A. Pease e M. Ellsworth. Linking framenet to the suggested upper merged ontology. In *Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS'06)*, pp. 289–300, Baltimore, USA, 2006a.

J. Scheffczyk, C.F. Baker e S. Narayanan. Ontology–based reasoning about lexical resources. In *Proceedings of the Workshop OntoLex 2006*, Genova, Italia, 2006b.

T. Schmidt. The kicktionary: Combining corpus linguistics and lexical semantics for a multilingual football dictionary. In E. Lavric, G. Pisek, A. Skinner e W. Stadler, editori, *The Linguistics of Football (Language in Performance 38)*, pp. 11–23. Tubingen, Gunter Narr, 2008.

L. Shi e R. Mihalcea. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing*, pp. 100–111, 2005.

R. Simone. testo parlato e testo scritto. In M. de las Nieves Muniz Muniz, editore, *La costruzione del testo in italiano. Sistemi costruttivi e testi costruiti*, pp. 23–61. Firenze, Franco Casati, 1996.

B. Smith e C. Fellbaum. Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004.

P.L. Spinosa, G. Giardiello, M. Cherubini, S. Marchi, G. Venturi e S. Montemagni. Nlp–based metadata extraction for legal text consolidation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pp. 40–49, Barcelona, Spain, 2009.

C. Subirats e M.R.L. Petruck. Surprise: Spanish framenet. In *Proceedings of the XVII International Congress of Linguists. Workshop on Frame Semantics*, Prague, Czech Republic, 2003.

G. Tarello. Orientamenti analitico–linguistici e teoria dell'interpretazione giuridica. In U. Scarpelli, editore, *Diritto e analisi del linguaggio*, pp. 375–395. Milano, Edizioni di Comunità, 1976.

D. Tiscornia. L'utilizzo di modelli della conoscenza nella legimatica. In C. Biagioli, P. Mercatali e G. Sartor, editori, *Legimatica. Informatica per legiferare*, pp. 313–338. Napoli, ESI, 1995.

D. Tiscornia. The lois project: Lexical ontologies for legal information sharing. In *Proceedings of the V Legislative XML Workshop*, pp. 189–204, San Domenico di Fiesole, Italia, 2007.

S. Tonelli. *Semi–automatic techniques for extending the FrameNet lexical database to new languages*. Ph.D. thesis, Università di Venezia, Dipartimento di Scienze del Linguaggio, 2010.

S. Uematsu, J-D. Kim e J. Tsujii. Bridging the gap between domain–oriented and linguistically–oriented semantics. In *Proceedings of the Workshop on BioNLP*, pp. 162–170, Boulder, Colorado, 2009.

R. van Kralingen. A conceptual frame-based ontology for the law. In *Proceedings of the First International Workshop on Legal Ontologies*, pp. 6–17, 1997.

R. van Kralingen, E. Oskamp e E. Reurings. Norm frames in the representation of laws. In *Proceedings of Legal Knowledge and Information Systems (JURIX) Conference*, pp. 11–21, 1993.

G. Venturi. Legal language and legal knowledge management applications. In W. Peters E. Francesconi, S. Montemagni e D. Tiscornia, editori, *Semantic Processing of Legal Texts*, LNCS, vol. 6036, pp. 3–26. Springer–Verlag, Berlin Heidelberg, 2010.

G. Venturi. Semantic annotation of italian legal texts: a framenet-based approach. In K. Ohara e K. Nikiforidou, editori, *Constructions and Frames*, Special issue, pp. 46–79. John Benjamins Company, 2011.

J. Visconti. A modular approach to legal translation. In G. Grewendorf e M. Rathert, editori, *Formal Linguistics and Law*, pp. 401–426. Mouton de Gruyter, 2009.

J. Visconti, editore. *Lingua e Diritto: Livelli Di Analisi*. LED – Edizioni Universitarie di Lettere Economia Diritto, 2010.

P. Vossen. Eurowordnet: Building a multilingual database with wordnets for several european languages. In *The ELRA Newsletter*, volume 3(1), pp. 7–10, 1998.

S. Walter. Definition extraction from court decisions using computational linguistic technology. In G. Grewendorf e M. Rathert, editori, *Formal Linguistics and Law*, pp. 183–224. Mouton de Gruyter, 2009.

T. Wattarujeekrit, P. Shah e N. Collier. Pasbio: predicate–argument structures for event extraction in molecular biology. In *BMC BioInformatics*, pp. 1–155, 2004.

A. Wyner. *Violations and Fulfillment in the Formal Representation of Contracts*. Ph.D. thesis, King's College London, 2008.

A. Wyner. Towards annotating and extracting textual legal case elements. In *Proceedings of the IV Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2010)*, pp. 9–18, Fiesole, Italia, 2010.

A. Wyner e W. Peters. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In *Proceedings of Legal Knowledge and Information Systems (JURIX) Conference*, pp. 127–136, Liverpool, United Kingdom, 2010a. IOS Press.

A. Wyner e W. Peters. Towards annotating and extracting textual legal case factors. In *Proceedings of the Language Resources and Evaluation Conference (LREC2010), Workshop on Semantic Processing of Legal Texts (SPLeT 2010)*, pp. 36–45, La Valletta, Malta, 2010b.

A. Wyner e T. van Engers. From argument in natural language to formalised argumentation: Components, prospects and problems. In *Proceedings of the Worskhop on Natural Language Engineering of Legal Argumentation (NaLEA2009)*, Barcelona, Spain, 2009.

G. Zaccaria. Testo, contesto e linguaggi settoriali nell'interpretazione giuridica. In P. Mariani Biagini, editore, *La lingua, la legge, la professione forense. Atti del convegno Accademia della Crusca (Firenze, 31 gennaio–1 febbraio 2002*, pp. 89–102. Milano, Giuffrè, 2003.

E. Zuanelli, editore. *Il diritto all'informazione in Italia*. Roma, Presidenza del Consiglio dei Ministri. Dipartimento per l'informazione e l'editoria, 1990.