In *Studi Italiani di Linguistica Teorica e Applicata* (SILTA)
Anno XLII, Numero 1, 2013, pp. 145-172

Simonetta Montemagni
(Istituto di Linguistica Computazionale "Antonio Zampolli" - ILC-CNR)

Tecnologie linguistico-computazionali e monitoraggio della lingua italiana

1 INTRODUZIONE

In una riflessione su dove stia andando l'italiano del terzo millennio, è legittimo chiedersi se e in che misura le tecnologie linguistico-computazionali possano essere di aiuto nel monitoraggio della lingua italiana nelle sue varietà diamesiche, diafasiche e diastratiche, nonché sull'asse diacronico. L'obiettivo del presente contributo consiste nel fornire una risposta, sebbene preliminare, a questo interrogativo, primariamente sul versante metodologico. In particolare, si vuole mostrare che mediante il ricorso a tecnologie linguistico-computazionali è oggi possibile monitorare un ampio spettro di tratti, che spaziano tra i diversi livelli di descrizione linguistica (primariamente, lessico, morfo-sintassi e sintassi), in relazione a corpora di sempre più vaste dimensioni. Questo rappresenta un cambio fondamentale nello studio della variazione linguistica, in particolare della lingua italiana, fino a oggi basato su corpora di dimensioni relativamente ridotte e tipicamente condotto mediante un'analisi (semi-)manuale del testo. Come vedremo, l'uso di vasti corpora testuali combinato con il ricorso a tecnologie linguistico-computazionali per l'analisi e il monitoraggio linguistico rendono oggi possibili analisi sempre più accurate e affidabili, che coprono aspetti della struttura linguistica rimasti fino a ora inesplorati in quanto difficilmente attingibili mediante un'analisi manuale del testo.

L'intuizione di partenza riguardante il "potere diagnostico" delle tecnologie linguistico-computazionali in compiti di monitoraggio linguistico trova conferma in un recente filone di studi avviato a livello internazionale all'interno del quale analisi linguistiche generate da strumenti di trattamento automatico del linguaggio sono usate, ad esempio, per:

- monitorare lo sviluppo della sintassi nel linguaggio infantile (Sagae *et al.*, 2005; Lu, 2008);
- identificare deficit cognitivi attraverso misure di complessità sintattica (Roark *et al.*, 2007);
- misurare la leggibilità di testi per studenti di L1 e L2 (Heilman *et al.*, 2007; Collins-Thompson, 2005);
- monitorare la capacità di lettura come componente centrale della competenza linguistica (Schwarm *et al.*, 2005; Petersen *et al.*, 2009).

Sulla scia di questi studi, abbiamo condotto primi esperimenti finalizzati al monitoraggio della lingua italiana nelle sue varietà d'uso. Nell'indagare le potenzialità di tali tecnologie nel monitoraggio della lingua italiana a partire dall'analisi automatica di corpora rappresentativi di diverse varietà di lingua e generi testuali, questo studio si focalizza sulla definizione di una metodologia che possa essere utilmente sfruttata sia

sul versante teorico sia in contesti applicativi (ad esempio per il monitoraggio delle competenze linguistiche in ambito scolastico, cfr. Dell'Orletta e Montemagni, 2012).

Le tecnologie linguistico-computazionali utilizzate a tal fine sono costituite da una piattaforma ormai consolidata e ampiamente sperimentata di metodi e strumenti per il trattamento automatico dell'italiano sviluppati presso l'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) che ha ricevuto ampia validazione sia nell'ambito di progetti di ricerca volti all'estrazione di informazione linguistica da corpora testuali, sia in progetti di carattere applicativo finalizzati all'estrazione di conoscenza di dominio. Il monitoraggio ha riguardato sia il lessico sia aspetti della struttura linguistica (morfo-sintattica e sintattica). Nella tipologia di parametri di monitoraggio indagati, l'aspetto di maggiore novità riguarda quelli basati su "microprelievi" effettuati sul testo arricchito con informazione morfo-sintattica e sintattica che, per quanto includa un inevitabile margine di errore, se appropriatamente esplorato rende possibile l'indagine di aspetti della struttura linguistica altrimenti difficilmente investigabili: è su questi aspetti che si incentra il presente contributo.

In quanto segue, la metodologia di monitoraggio proposta verrà illustrata in dettaglio: nella sezione 2 sono fornite le nozioni di base relative alle tecnologie linguistico-computazionali utilizzate; la sezione 3 descrive i corpora su cui questo studio si è basato, mentre la sezione 4 discute la tipologia di parametri di monitoraggio identificabili a partire da testi arricchiti con annotazione linguistica automatica. La sezione 5 fornisce una rivisitazione di un caso di studio ampiamente dibattuto nella letteratura linguistica – ovvero la distribuzione di nomi e verbi – alla luce dell'evidenza emersa dall'analisi automatica di corpora di vaste dimensioni rappresentativi di diverse varietà d'uso della lingua italiana.

2 LE TECNOLOGIE LINGUISTICO-COMPUTAZIONALI

Le tecnologie linguistico-computazionali permettono di accedere al contenuto informativo dei testi attraverso l'individuazione della struttura linguistica sottostante e la sua rappresentazione esplicita. L'identificazione della struttura linguistica del testo avviene tipicamente in modo incrementale, attraverso analisi linguistiche a livelli di complessità crescente: "tokenizzazione", ovvero segmentazione del testo in parole ortografiche (o "tokens"); analisi morfo-sintattica e lemmatizzazione del testo "tokenizzato"; analisi della struttura sintattica della frase in termini di relazioni di dipendenza. La Tabella 1 esemplifica il risultato di questo processo di analisi incrementale: ogni riga corrisponde a un'occorrenza di forma di parola ("token"), mentre le colonne specificano le proprietà di questa forma ai diversi livelli di analisi. All'interno di una frase, ogni forma è univocamente identificata da un numero progressivo (colonna 1). Al livello di annotazione morfo-sintattica, a ogni "token" del testo viene associata informazione relativa alla categoria grammaticale che la parola ha nel contesto specifico (in colonna 4, V = verbo; R = articolo, S = sostantivo, A = aggettivo, E = preposizione; in colonna 5, eventuali sottocategorie come ad es. EA = preposizione articolata, RD-RI = articolo definito-indefinito). Tale informazione è integrata da ulteriori specificazioni morfologiche (in colonna 6, riguardanti ad es. categorie flessionali come persona, genere, numero, ecc.) e il relativo esponente lessicale o lemma (colonna 3).

		Lemmatizzazione	Annotazione morfo-sintattica			Annotazione a dipendenze	
Id	Forma	Lemma	CaGra1	CaGra2	Tratti	Testa	Tipo di relazione
1	Le	il	R	RD	num=p gen=f	2	det
2	tecnologie	tecnologia	S	S	num=p gen=f	4	subj
3	linguistiche	linguistico	A	A	num=p gen=f	2	mod
4	rappresentano	rappresentare	V	V	num=p per=3 mod=i ten=p	0	ROOT
5	un	un	R	RI	num=s gen=m	6	det
6	ausilio	ausilio	S	S	num=s gen=m	4	obj
7	importante	importante	A	A	num=s gen=n	6	mod
8	per	per	E	E	_	6	comp
9	il	il	R	RD	num=s gen=m	10	det
10	monitoraggio	monitoraggio	S	S	num=s gen=m	8	prep
11	della	di	E	EA	num=s gen=f	10	comp
12	lingua	lingua	S	S	num=s gen=f	11	prep
13	italiana	italiano	A	A	num=s gen=f	12	mod
14	.	.	F	FS	_	4	punc

Tabella 1: Esempio di rappresentazione tabellare del testo annotato linguisticamente

Il livello di annotazione sintattica, rappresentato nelle ultime due colonne della tabella, fornisce invece una descrizione della frase in termini di relazioni binarie di dipendenza tra parole (tipicamente, relazioni binarie asimmetriche tra una testa e un dipendente, come "soggetto", "oggetto diretto", "modificatore", etc.).[1] Più concretamente, per ogni parola la colonna 7 riporta l'identificatore univoco della forma che costituisce la testa da cui dipende (0 per il verbo della proposizione principale, assunto come radice dell'albero sintattico), mentre la colonna 8 specifica il tipo di dipendenza. Dalla Tabella 1, si evince ad esempio che la parola *tecnologie* costituisce il soggetto (subj) di *rappresentano* (Id=4) e *ausilio* l'oggetto (obj); che *linguistiche* è un modificatore (mod) della testa *tecnologie* (Id=2), così come *importante* lo è rispetto ad *ausilio* (Id=6). La Figura 1 fornisce una rappresentazione grafica dell'albero di dipendenze sintattiche in Tabella 1, all'interno della quale gli archi marcano la dipendenza sintattica che lega la testa al dipendente.

[1] Per maggiori dettagli sul tipo di rappresentazione sintattica adottata, si rinvia il lettore interessato a Lenci *et al.* (2009).

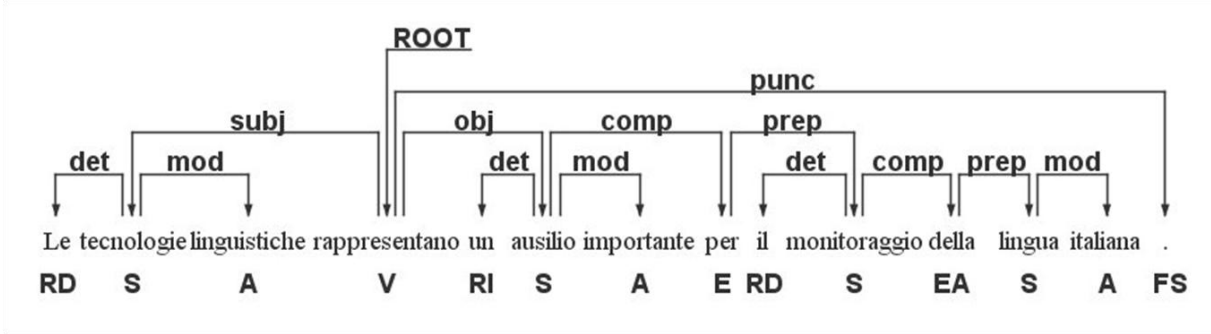

Figura 1: Rappresentazione grafica dell'annotazione linguistica dell'esempio in Tabella 1

Un testo arricchito con informazioni di questo tipo diventa il punto di partenza per ulteriori elaborazioni automatiche, in particolare per l'identificazione di una vasta tipologia di parametri che possono essere utilmente sfruttati in compiti di monitoraggio linguistico.

Oggi, nel campo della linguistica computazionale, lo stato dell'arte nei compiti di annotazione linguistica è rappresentato da sistemi basati su algoritmi di apprendimento automatico supervisionato. Il compito di "annotazione linguistica" viene modellato come un compito di classificazione probabilistica: a ogni passo di computazione il sistema sceglie l'annotazione più probabile data la parola in input, i suoi tratti descrittivi, il contesto e le annotazioni linguistiche già identificate. A partire da un corpus di addestramento, annotato con informazione morfo-sintattica e sintattica, viene costruito un modello probabilistico per l'annotazione linguistica del testo.

Gli strumenti software alla base della metodologia di monitoraggio linguistico proposta in questo lavoro rappresentano lo "stato dell'arte" per la lingua italiana, in quanto sono risultati gli strumenti più precisi e affidabili per l'annotazione morfo-sintattica e sintattica a dipendenze nella campagna di valutazione di strumenti per l'analisi automatica dell'italiano, EVALITA-2009 (Evalita, 2009). Per quanto riguarda l'annotazione morfo-sintattica, lo strumento utilizzato (Dell'Orletta, 2009) presenta un'accuratezza[2] del 96,34% nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati (senza utilizzare alcuna risorsa lessicale di riferimento). Per quanto riguarda l'analisi a dipendenze, abbiamo utilizzato DeSR (Attardi *et al.*, 2009). Data la maggiore complessità del compito di analisi sintattica, per calcolare l'accuratezza di un analizzatore a dipendenze vengono usate diverse metriche. Tra queste, ai fini del presente studio vale la pena menzionare:

i. "Labelled Attachment Score" (LAS), ovvero la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda sia la testa sintattica sia il ruolo (tipo di relazione di dipendenza) svolto in relazione ad essa;
ii. "Unlabelled Attachment Score" (UAS), ovvero la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione della testa sintattica.

[2] L'accuratezza è stata calcolata come il rapporto tra il numero di token classificati correttamente e il numero totale di token analizzati.

DeSR raggiunge livelli di LAS e UAS in linea con lo stato dell'arte dell'analisi a dipendenze, pari a 83,38% e 87,71% rispettivamente; ovvero, DeSR è molto affidabile nel ricostruire le relazioni di dipendenza che collegano le parole della frase (UAS), mentre più problematica appare l'identificazione simultanea del tipo di dipendenza e della testa sintattica (LAS). Come vedremo nella sezione 4, molti dei "micro-prelievi" alla base della metodologia di monitoraggio proposta si basano o sull'identificazione della relazione di dipendenza che lega una testa a un dipendente: a questo livello, abbiamo visto che l'affidabilità dell'analizzatore è sensibilmente maggiore.

Possiamo dunque concludere che per quanto i risultati dell'annotazione linguistica automatica includano inevitabilmente un margine di errore, che abbiamo visto variare a seconda del livello e del tipo di informazione linguistica considerata, se appropriatamente esplorati possono fornire indicazioni affidabili nella ricostruzione del profilo linguistico di un testo.

3 I CORPORA USATI

In linea di principio, uno studio di monitoraggio linguistico dovrebbe basarsi su una sorta di "monitor corpus" secondo l'accezione sinclairiana (Sinclair, 1991), ovvero una collezione "aperta" di testi che muta nel tempo secondo criteri prestabiliti per poter tener traccia dei processi di variazione linguistica tra diverse varietà d'uso, generi testuali così come in una prospettiva diacronica. Attraverso la nozione di "monitor corpus" (o corpus di monitoraggio) è possibile monitorare le dinamiche del lessico e più in generale della lingua a diversi livelli di descrizione linguistica.

Etichetta	Corpus	Fonte	Parole	Periodi	N° medio di parole per periodo
Rep	*Repubblica 2002-2005*	Corpus di Lingua Italiana Contemporanea (CLIC – ILC, Marinelli *et al.*, 2003)	10.306.624	444.807	23,17
Narr	*Narrativa 1974-1989*	Corpus di Lingua Italiana Contemporanea (CLIC – ILC, Marinelli *et al.*, 2003)	1.242.802	55.947	22,21
RaccFant	*Racconti fantastici 2007-2008*	Marinelli *et al*. (2008)	25.492	1.149	22,19
Suss	*Sussidiari delle scuole elementari 2010*	Progetto CNR "Migrazioni"	47.993	2.521	19,04
2Par	*Due Parole 2001-2006*	http://www.dueparole.it/	73.031	6.063	12,05

Etichetta	Corpus	Fonte	Parole	Periodi	N° medio di parole per periodo
Giur	*Corpus legislativo ambientale 1997-2005*	Venturi (2006)	1.645.286	72.501	22,69
Parlato	*Parlato*	C-ORAL-ROM http://lablita.dit.unifi.it/coralrom	340.275	37.078	9,18
Parlato formale	*Parlato formale*	Parlato formale primariamente di tipo monologico in contesto naturale o in contesti media	189.383	17.560	10,78
Parlato informale	*Parlato informale*	Parlato naturale in forma dialogica e monologica in ambito familiare	150.892	19.518	7,73

Tabella 2: I corpora di monitoraggio selezionati

Non disponendo al momento di tale corpus, abbiamo condotto questo studio preliminare, primariamente di carattere metodologico, su corpora di varia natura (riportati in Tabella 2) rappresentativi di diversi generi testuali e varietà di lingua. Per la lingua scritta sono stati selezionati: prosa giornalistica (Rep); narrativa (Narr); testi legislativi (Giur); racconti fantastici (RaccFant), costituiti da testi che descrivono situazioni immaginarie che sono stati prodotti nell'ambito di uno studio volto a valutare possibili differenze linguistiche in soggetti caratterizzati da diversi gradi di suscettibilità ipnotica. A questi corpora, caratterizzati da diversi livelli di complessità linguistica, sono stati affiancati due corpora rappresentativi di un linguaggio semplificato, costituiti dal periodico di "facile lettura" *Due Parole* (2Par) indirizzato a persone con deficit cognitivi o caratterizzate da un basso livello di alfabetizzazione, e da sussidiari della scuola primaria (Suss).

Come corpus di parlato spontaneo abbiamo utilizzato C-ORAL-ROM Italia, il sottocorpus italiano di C-ORAL-ROM, un corpus delle principali lingue parlate romanze (italiano, francese, spagnolo e portoghese) realizzato da un consorzio di importanti istituzioni europee coordinato da Manuela Cresti (per maggiori dettagli sulla risorsa cfr. Cresti e Moneglia, 2005). Ai fini delle nostre analisi, abbiamo sfruttato l'articolazione interna del corpus basata sulla distinzione tra parlato formale e informale (Moneglia, 2004), suddividendolo in due partizioni rappresentative dei due generi di parlato: il parlato informale e il parlato formale.

Un corpus così variegato offre la possibilità di monitorare differenze e similarità linguistiche a diversi livelli, ad esempio: all'interno dell'opposizione testi informativi (costituititi in questo studio da Rep, 2Par e Giur) vs testi di scrittura creativa (Narr, RaccFant e Suss); tra testi complessi come la prosa giornalistica (Rep) e i testi legislativi (Giur) e testi semplificati (qui congiuntamente rappresentati da Suss e 2Par);

infine, la presenza di un corpus di parlato spontaneo rende anche possibile l'indagine del rapporto tra parlato e scritto.

4 MONITORAGGIO LINGUISTICO: PRINCIPI E METODI

Fino ad oggi, gli studi su varietà di lingua e generi testuali così come sull'opposizione parlato vs scritto per la lingua italiana si sono tipicamente basati su evidenza tratta da lessici di frequenza costruiti a partire da corpora il cui processo di lemmatizzazione e annotazione morfo-sintattica è stato condotto in modo manuale o semi-automatico. Per la lingua scritta, si va dal *Lessico di frequenza della lingua italiana contemporanea* (LIF, 1972), al *Vocabolario Elettronico della Lingua Italiana (*VELI, 1989), a CoLFIS (Bertinetto *et al.*, 2005), per arrivare a lessici che potremmo definire "settoriali" quali il *Lessico Elementare* (LE, 1994) o il lessico di frequenza basato su quattro annate del giornale *Due Parole* (Piemontese, 1996). Per il parlato, si ha il *Lessico di Frequenza dell'Italiano Parlato* (LIP, 1993), e più recentemente C-ORAL-ROM (Cresti e Moneglia, 2005). A parte C-ORAL-ROM in cui l'annotazione morfo-sintattica è stata condotta in modo automatico (Panunzi *et al.*, 2004), in tutti gli altri casi al processo automatico di lemmatizzazione e disambiguazione morfo-sintattica delle parole dei corpora ha fatto seguito una fase di revisione manuale. E' su risorse di questo tipo che sono basati, ad esempio, gli studi di Voghera (2004, 2005) sulla distribuzione delle parti del discorso (o categorie morfo-sintattiche) nel parlato e nello scritto così come tra diversi tipi di testi, o quelli di Cresti (2005) sulle strategie lessicali tipiche del parlato rispetto alla lingua scritta: con una differenza sostanziale nell'ottica del presente studio, costituita dal fatto che mentre gli studi di Voghera sono basati su corpora la cui annotazione morfo-sintattica è stata oggetto di revisione manuale, lo studio di Cresti si basa su un corpus annotato automaticamente.

Va tuttavia considerato che i dati tratti da risorse di questo tipo "concernono unicamente la lista dei lemmi e non illustrano la funzione sintattico-semantica che essi assumono, nella predicazione, nei vari contesti frastici" (D'Agostino, 1998:19). Per condurre uno studio della variazione linguistica tra varietà d'uso e generi testuali al livello della sintassi sono necessarie risorse che includano un livello di annotazione sintattica. A nostra conoscenza, nel panorama italiano l'unico corpus che è stato concepito a supporto dello studio delle differenze a livello sintattico tra diverse varietà d'uso della lingua italiana è rappresentato dal Corpus Penelope (www.parlaritaliano.it), una risorsa di dimensioni contenute - poco più di 30.000 parole – che è stata costruita per lo studio di "sistematicità sintattiche": nei suoi venticinque anni di vita, tale corpus ha costituito la campionatura di riferimento per studi e ricerche di tipo sintattico, per tener traccia di tendenze e trasformazioni della lingua in prospettiva sia sincronica sia diacronica e per l'analisi di specificità di linguaggi settoriali (ad esempio, il linguaggio politico o il linguaggio giovanile). In questa prospettiva si inquadrano gli studi di Policarpi e Rombi sulle tendenze nella sintassi dell'italiano contemporaneo (Rombi e Policarpi, 1985; Policarpi e Rombi, 2005). In linea di principio, una possibile alternativa al corpus Penelope potrebbe essere rappresentata da risorse quali le Treebank sintattiche per la lingua italiana (ISST, Montemagni *et al.*, 2003; TUT, Bosco *et al.*, 2000; VIT, Delmonte *et al.*, 2007) che tuttavia non possono essere utilizzate ai fini dello studio della variazione linguistica tra varietà di lingua e generi testuali in

quanto la composizione interna dei corpora annotati non era stata definita come rappresentativa di diversi generi testuali.

Per altre lingue, la situazione è differente. Fino dalla metà degli anni Ottanta, prima per la lingua inglese (Biber, 1988) e poi anche per altre lingue quali il somalo, il coreano, il taiwanese (Biber, 1995) e anche lo spagnolo (Biber *et al.*, 2006) si annoverano studi su "register variation" basati su corpora di maggiori dimensioni, dell'ordine di milioni di parole, con l'analisi linguistica del testo condotta in modo completamente automatico; tali studi sono condotti ricorrendo a tecniche di statistica multivariata che rendono possibile l'identificazione dell'insieme dei tratti caratterizzanti una varietà di lingua rispetto a un'altra. Ai fini del presente studio, questo filone di letteratura, che ha origine nei lavori pioneristici di Biber per la lingua inglese, riveste un ruolo particolarmente importante in quanto mostra che usando corpora di vaste dimensioni e strumenti di analisi linguistica automatica del testo è possibile condurre un monitoraggio linguistico ad ampio spettro, che coinvolge una varia ed estesa tipologia di parametri riguardanti i diversi livelli di descrizione linguistica. Tutti questi studi si basano su un livello di annotazione morfo-sintattica "potenziato" con regole ad hoc operanti sulla sequenza delle categorie grammaticali e sui lemmi per l'identificazione di particolari costruzioni sintattiche e strutture semantiche.

Il presente contributo si colloca all'interno di questo filone di studi e si concentra sulla tipologia di parametri che possono essere oggetto di monitoraggio linguistico facendo ricorso a tecnologie linguistico-computazionali, che forniscono un livello di analisi più avanzato sia rispetto alla tradizione di studi sulla lingua italiana, sia rispetto alla letteratura corrente sulla "register variation". Tali tecnologie cominciano infatti a essere mature per poter essere sfruttate in compiti di monitoraggio linguistico, con un'accuratezza che, seppur decrescente attraverso i diversi livelli di annotazione, è sempre più che accettabile. Ciò implica che il monitoraggio linguistico oggi può essere condotto in relazione a corpora di vaste dimensioni e basarsi su informazioni della struttura sintattica che fino ad ora sembravano essere inattingibili se non attraverso un accurato lavoro manuale che per sua natura non poteva che essere circoscritto a limitate porzioni di testo e a ristretti repertori di tratti linguistici.

In quanto segue, cercheremo di delineare la tipologia di parametri che possono essere monitorati a partire dai diversi livelli di annotazione linguistica, in particolare morfo-sintattica e sintattica a dipendenze.

4.1 Livelli di annotazione e parametri di monitoraggio linguistico

Partiamo dal livello di annotazione morfo-sintattica: attraverso la "misura" delle categorie morfo-sintattiche è possibile rintracciare differenze e somiglianze tra diversi generi testuali così come tra scritto e parlato. La Tabella 3 registra la distribuzione percentuale delle categorie morfo-sintattiche nei corpora di lingua scritta considerati, mentre l'ultima colonna ne riporta l'occorrenza media.

	Rep	Narr	RaccFant	Suss	2Par	Giur	Media
Aggettivi	6,51	7,09	7,41	6,42	5,90	8,40	6,96
Articoli	8,20	7,64	8,33	9,41	10,34	6,43	8,39
Avverbi	5,13	5,42	7,01	6,08	3,44	1,53	4,77
Congiunzioni	4,06	4,85	5,32	5,38	3,83	4,15	4,60

	Rep	Narr	RaccFant	Suss	2Par	Giur	Media
Determinanti	0,90	1,13	1,38	1,00	1,66	0,47	1,09
Interiezioni	0,02	0,05	0,05	0,08	0,00	0,00	0,03
Numerali	1,98	1,31	0,46	0,63	2,63	5,94	2,16
Predeterminanti	0,13	0,12	0,13	0,21	0,32	0,08	0,17
Preposizioni	15,47	13,81	13,70	12,65	15,43	20,71	15,30
Pronomi	4,27	5,99	8,13	6,06	2,18	2,02	4,77
Sostantivi	26,43	23,33	19,76	21,09	29,68	29,60	24,98
Verbi	13,09	14,25	16,98	16,52	13,52	8,77	13,85

Tabella 3 – Distribuzione delle categorie morfo-sintattiche nello scritto

Ad un'analisi dei dati nella tabella, si osserva che le categorie maggiormente oscillanti tra i diversi generi testuali sono rappresentate da numerali, preposizioni, pronomi, nomi e verbi (caratterizzate, tutte, da una deviazione standard > 2). Sulle ultime due classi ritorneremo nella sezione 5. Qui vale la pena segnalare che un'alta ricorrenza di numerali e preposizioni sembra contraddistinguere i testi giuridici, mentre i pronomi rappresentano un tratto caratterizzante di testi di scrittura creativa.

La Tabella 4 riporta la distribuzione dello stesso insieme di categorie nel parlato, distinguendo tra parlato formale e informale (seconda e terza colonna). In questo caso, le categorie che presentano una maggiore oscillazione nell'uso sono gli aggettivi, le interiezioni, le preposizioni, i pronomi e i sostantivi (per questi ultimi, in relazione ai verbi, si rinvia alla sezione 5). Confrontando i dati delle Tabelle 3 e 4, in particolare la media per lo scritto e il dato globale per la lingua parlata (prima colonna della Tabella 4), si nota che le maggiori differenze tra scritto e parlato riguardano la distribuzione di avverbi, preposizioni, pronomi, sostantivi e verbi.

	Parlato	Parlato formale	Parlato informale
Aggettivi	4,63	5,38	3,68
Articoli	7,13	7,32	6,89
Avverbi	10,77	10,34	11,32
Congiunzioni	5,69	5,64	5,74
Determinanti	1,58	1,76	1,36
Interiezioni	1,33	0,93	1,83
Numerali	1,32	1,44	1,18
Predeterminanti	0,20	0,19	0,21
Preposizioni	10,50	12,36	8,17
Pronomi	9,41	8,41	10,67
Sostantivi	17,91	19,18	16,31
Verbi	17,27	17,10	17,49

Tabella 4 – Distribuzione delle categorie morfo-sintattiche nel parlato

I dati estratti dai corpora di monitoraggio considerati sono in linea con la letteratura sull'argomento, per menzionarne alcuni LIP (1993), Voghera (2004), Cresti (2005). Ciò mostra che il dato attinto dal corpus annotato automaticamente riflette in modo affidabile tendenze già note, calcolate a partire da corpora la cui annotazione (a parte il caso di Cresti) è stata oggetto di revisione manuale.

Questa coincidenza di risultati ci incoraggia a proseguire con nuovi tipi di misure che non hanno equivalenti nella letteratura precedente. Innanzitutto, la misura delle categorie morfo-sintattiche può essere condotta anche in relazione a sotto-classi quali, ad esempio, congiunzioni coordinanti vs subordinanti. Nella Tabella 3 abbiamo osservato che la distribuzione della categoria delle congiunzioni per lo scritto varia tra 3.83% nel caso di 2Par al più del 5% nel caso di Suss (5.38%) e RaccFant (5.32%). Ma se andiamo a vedere come le due sotto-categorie grammaticali delle congiunzioni coordinanti e subordinanti si distribuiscono all'interno dei diversi corpora (Tabella 5), si osservano differenze significative non desumibili dal dato aggregato:

	Rep	Narr	RaccFant	Suss	2Par	Giur
Cong. coordinanti	70,77	69,08	71,46	70,47	79,25	87,54
Cong. subordinanti	29,23	30,92	28,54	29,53	20,75	12,46

Tabella 5 – Ripartizione interna della classe delle congiunzioni in coordinanti vs subordinanti

Se assumiamo la percentuale di congiunzioni subordinanti quale indicatore della proporzione di costruzioni ipotattiche all'interno del testo, possiamo osservare che i corpora Giur e 2Par fanno un limitato ricorso a tali tipi di costruzioni: ciò è spesso associato a una maggiore "leggerezza" sintattica. In tutti gli altri casi (ovvero, Rep, Narr, RaccFant e Suss), la proporzione delle congiunzioni subordinanti si attesta attorno al 30%. Questo dato fornisce una prima e soprattutto molto approssimativa indicazione del rapporto tra costruzioni paratattiche e ipotattiche all'interno dei corpora selezionati. Va infatti tenuto presente che se da un lato le congiunzioni subordinanti sono tipicamente associate a clausole subordinate, nel caso delle congiunzioni coordinanti va considerato il fatto che possono riguardare diverse categorie grammaticali, non solo i verbi.[3]

A partire dalla distribuzione delle singole categorie morfo-sintattiche, è possibile mettere in relazione la frequenza di occorrenza di certe categorie grammaticali rispetto ad altre: esempio, misurare la "densità lessicale" calcolata come la proporzione delle parole semanticamente "piene" (ovvero, nomi, aggettivi, verbi e avverbi) rispetto al totale delle occorrenze (Dell'Orletta e Montemagni, 2012), oppure il rapporto tra diverse categorie morfo-sintattiche, ad esempio il rapporto tra nomi e verbi (cfr. infra).

Come già anticipato, esistono tuttavia aspetti della struttura sintattica che non possono essere ricavati a partire dalla distribuzione delle categorie morfo-sintattiche nel testo. Al massimo, a partire dalla sequenza delle categorie morfo-sintattiche è possibile ricostruire la presenza di strutture particolari: questo è l'approccio seguito da Biber nei suoi studi su "register variation", che opera sul testo annotato morfo-sintatticamente

[3] Per poter indagare sui diversi tipi di strutture coordinate al fine di estrapolare solo quelle di tipo verbale è necessario partire da un'annotazione sintattica a dipendenze: solo in questo modo è possibile ricostruire quale sia la tipologia degli elementi coinvolti nella costruzione coordinata.

all'interno del quale vengono ricercate sequenze di categorie grammaticali e/o lemmi per l'identificazione automatica di un ampio spettro di tratti riguardanti diversi aspetti della struttura linguistica. Ad esempio, la presenza di una clausola infinitiva viene rintracciata attraverso la ricerca dello schema "*to* + (ADV) + VB" da interpretarsi come particella *to*, seguita opzionalmente da un avverbio, a sua volta seguita da un verbo nella sua forma base (Biber, 1988:232); i complementi preposizionali sono identificati in base alla presenza di preposizioni, a partire da una lista che include solo elementi non ambigui, ovvero non suscettibili di altra interpretazione morfo-sintattica, quale ad esempio avverbio o congiunzione (Biber, 1988:236).

Per quanto questo tipo di informazione rappresenti un passo in avanti rispetto all'analisi della distribuzione delle categorie morfo-sintattiche considerate al di fuori del contesto di occorrenza, non è tuttavia sufficiente a ricavare indicazioni precise in merito alla struttura sintattica complessiva sottostante al testo oggetto dell'analisi. Oltre a rilevare la sua presenza, di una clausola infinitiva potremmo voler sapere la categoria morfo-sintattica della testa da cui dipende (ad esempio, verbo vs nome, come in *ho deciso di partire* vs *la decisione di partire*), oppure il suo livello di incassamento, ovvero se dipenda direttamente dalla radice verbale della frase oppure se rappresenti una reggenza di secondo, terzo o quarto grado (ad esempio, *Giovanni ha deciso di partire* vs *Maria ha sentito che Giovanni avrebbe affermato di avere deciso di partire domani*). Di un complemento preposizionale può non bastare rilevare la sua presenza ma sarebbe utile – di nuovo – sapere da chi (verbo, nome o aggettivo) è governato, oppure se ricorra all'interno di un sintagma "leggero" o "pesante", e nel caso si tratti di un sintagma "pesante" fornisca una sorta di misura del suo peso. Questa tipologia di informazioni non è ricostruibile a partire dal testo annotato morfo-sintatticamente e neppure dalla distribuzione dei tipi di dipendenza all'interno del testo annotato, quanto dalle caratteristiche strutturali dell'albero sintattico.

Si consideri, in proposito, la distribuzione delle teste verbali nel corpus, che può essere inquadrata da prospettive diverse. Un dato elementare, ma già significativo, è costituito dal rapporto tra clausole e periodi, calcolato a partire dal numero di teste verbali rispetto al numero di periodi (cfr. infra): da un lato si ha il corpus Giur caratterizzato da una maggiore proporzione di periodi monoclausali, con una media di 1,64 clausole per periodo; dall'altro abbiamo Suss e RaccFant, contraddistinti da un valore medio di clausole per periodo sensibilmente più alto (3,37 e 2,71). Questo dato, però, non dice nulla su come le diverse clausole si rapportino l'una con l'altra all'interno del periodo: è possibile procedere in questa indagine andando a identificare, per ciascun corpus, la proporzione di principali e subordinate che può essere ricostruita a partire dal rapporto tra le radici verbali (corrispondenti alle frasi principali) e le clausole argomentali (ovvero sottocategorizzate dal verbo reggente) e quelle con valore temporale, causale, locativo, etc. In relazione a ciò, si osserva un andamento analogo in due corpora che sono agli antipodi, ovvero Giur e 2Par, che registrano entrambi una bassa percentuale di subordinate, che si attesta attorno al 25%; sull'altro versante si collocano gli altri corpora caratterizzati da una percentuale di subordinate che va dal 35% fino a un massimo di 43% nel caso di RaccFant. Tali dati (documentati in dettaglio in Dell'Orletta e Montemagni, 2012) sono in linea con quanto registrato in letteratura (Voghera 2001).

Può essere interessante integrare questa informazione con altri aspetti della struttura linguistica, quali ad esempio l'ordine relativo tra principale e subordinata, il

grado di incassamento della subordinata, il tipo di subordinata, etc. Per quanto riguarda l'ordine relativo della/e subordinata/e rispetto alla principale nei vari corpora, è interessante notare che nei RaccFant – un corpus particolarmente vicino alla lingua parlata – la maggiore ricorrenza di costruzioni subordinate è in un certo senso controbilanciata dal fatto che nella maggior parte dei casi (più del 94%) la subordinata segue la principale, un ordine che è riconosciuto di più facile elaborazione nella letteratura linguistica (cfr. Miller e Weinert, 1998). Negli altri casi, la percentuale di subordinate che seguono la principale è nettamente inferiore, attestandosi attorno all'87-88%.

Un altro aspetto rilevante riguarda i livelli di incassamento gerarchico: in presenza di più di una clausola subordinata all'interno dello stesso periodo, diventa cruciale ricostruire quale tipo di rapporto sussista tra di esse, ovvero se siano ricorsivamente incassate l'una all'interno dell'altra. Una prima e approssimativa misura dei livelli di incassamento gerarchico all'interno della struttura sintattica può essere ricostruita a partire dall'altezza massima dell'albero, che misura la massima distanza che intercorre tra una foglia (rappresenta da parole del testo senza dipendenti) e la radice dell'albero, espressa come numero di archi (ovvero relazioni di dipendenza) attraversati nel cammino foglia-radice. Questa misura può essere raffinata focalizzandosi su particolari tipi di sotto-alberi: ad esempio, sotto-alberi di clausole subordinate ricorsivamente incassate (cfr. Dell'Orletta e Montemagni, 2012), oppure sotto-alberi governati da una testa nominale (cfr. infra). A conclusione di questa breve esemplificazione di caratteristiche linguistiche che possono essere monitorate in relazione a testi annotati in modo automatico con informazione relativa alle dipendenze sintattiche, vale la pena fare notare che nelle misure esemplificate finora l'affidabilità dell'evidenza linguistica acquisita dai testi analizzati è riconducibile alla UAS (cfr. sezione 2), che abbiamo visto essere sensibilmente più alta della misura standard (LAS) di valutazione degli analizzatori sintattici a dipendenze.

Gli esempi riportati sopra non sono certo esaustivi di quanto è attingibile a partire da una rappresentazione a dipendenze, ma soltanto esemplificativi del valore aggiunto rappresentato da un livello di annotazione a dipendenze. Gli studi finora condotti sulla sintassi dell'italiano attraverso diversi generi testuali (anche in una prospettiva diacronica) hanno riguardato primariamente i rapporti tra clausole e periodi, oppure tra clausole principali e subordinate, e sono stati condotti su un corpus annotato in modo completamente manuale (Policarpi e Maggi, 2005) con inevitabili ripercussioni sia al livello della tipologia delle possibili analisi sia della rappresentatività dei risultati. D'altro canto, dalla letteratura linguistica, linguistico-computazionale e psicolinguistica si sa che vi sono parametri incentrati sulle caratteristiche strutturali dell'albero sintattico che rivestono un ruolo centrale nella valutazione della complessità di un testo. Ad esempio, come affermato da Yngve (1960), Frazier (1985) e Gibson (1998) da prospettive diverse, la "misura" della profondità dell'albero sintattico associato a una frase rappresenta un aspetto centrale nella valutazione della sua complessità. Un altro fattore di complessità ampiamente riconosciuto nella letteratura linguistica, psicolinguistica e linguistico-computazionale (cfr. Lin, 1996; Gibson, 1998) riguarda la "misura" della lunghezza delle relazioni di dipendenza, calcolata come la distanza (in parole) tra la testa e il dipendente. E' nostra convinzione che questi stessi parametri possano giocare un ruolo importante anche nel monitoraggio della variazione linguistica tra diversi generi testuali e/o varietà di lingua. Ovviamente, si tratta di ipotesi ad oggi

inesplorate in quanto non perseguibili senza analisi di tipo computazionale basate sulla struttura sintattica del testo.

Questa rapida e inevitabilmente limitata rassegna di quanto può essere estratto da testi arricchiti con annotazione linguistica multi-livello (in particolare, morfo-sintattica e sintattica) può non essere sufficiente a mostrare il potenziale di tale informazione nel monitoraggio linguistico di diverse varietà di lingua. Proponiamo dunque di seguito la rivisitazione di un caso di studio rispetto al quale esiste una letteratura consolidata sia in relazione all'italiano sia ad altre lingue: la distribuzione di nomi e verbi. E' nostra convinzione che in questo modo sia possibile valutare appieno il ruolo e l'impatto della tipologia di parametri illustrati in questa sezione nello studio della variazione linguistica.

5 NOMI VS VERBI E OLTRE

Che la frequenza dei nomi e dei verbi sia connessa a variazioni di tipo diamesico, diafasico e testuale rappresenta ad oggi un dato acquisito della letteratura linguistica e psicolinguistica. Come illustrato in Voghera (2005), la distribuzione dei nomi e dei verbi in corpora testuali offre spunti importanti per lo studio delle differenze tra parlato e scritto così come tra varietà di lingua e generi testuali.

Vediamo cosa emerge in relazione alla distribuzione dei nomi e dei verbi a partire dalle analisi automatiche condotte sui corpora di monitoraggio selezionati. La Tabella 6, nelle prime due righe, riporta i dati relativi alla distribuzione di nomi e verbi nella lingua scritta (questo dato, già riportato in Tabella 3, viene ripetuto di seguito per convenienza del lettore). Mentre le colonne 1-6 riportano la distribuzione di nomi e verbi in relazione a ciascun corpus analizzato, nell'ultima colonna viene riportata la frequenza media di occorrenza di nomi e verbi nei diversi generi testuali considerati. La terza riga della tabella fornisce una sintesi delle prime due righe, specificando il rapporto nomi/verbi ottenuto dividendo il numero dei nomi rilevati in ciascun corpus rispetto a quello dei verbi.

	Rep	Narr	RaccFant	Suss	2Par	Giur	Media
Nomi	26,43	23,33	19,76	21,09	29,68	29,60	24,98
Verbi	13,09	14,25	16,98	16,52	13,52	8,77	13,85
Rapporto nomi/verbi	2,02	1,64	1,16	1,28	2,20	3,37	1,80
Clausole x periodo	2,41	2,65	3,37	2,71	1,26	1,64	2,34
Nomi x clausola	2,55	1,96	1,30	1,48	2,84	4,09	2,37

Tabella 6: Parametri relativi alla distribuzione di nomi e verbi nella lingua scritta in diversi generi testuali basati sul testo annotato morfo-sintatticamente

Se da un lato i dati della tabella confermano che lo scritto è caratterizzato da una maggiore frequenza di occorrenza di nomi rispetto ai verbi, dall'altro si osservano variazioni significative tra i diversi corpora considerati: in particolare, si nota una sorta di divisione tra testi giornalistici, anche nella versione di "facile lettura", e legislativi da un lato e testi di scrittura creativa (narrativa, racconti fantastici e sussidiari) dall'altro. Nei primi, la frequenza di occorrenza dei nomi è maggiore rispetto alla media, mentre nei secondi sono piuttosto i verbi a ricorrere con valori più alti rispetto alla media (per i

nomi, in questo caso, si registrano valori al di sotto della media). Tale dato trova conferma nell'andamento del rapporto nomi/verbi, significativamente più alto della media nel caso della prima classe di testi, con valori nettamente inferiori nel caso della seconda classe.

La Tabella 7 registra la stessa tipologia di dati per la lingua parlata. I valori delle prime tre colonne si riferiscono all'output del livello di annotazione morfo-sintattica generato con uno strumento addestrato su corpora di lingua scritta (cfr. sezione 2). Tali dati sono in linea con quanto riportato in Cresti (2005) dove risulta che per la lingua italiana la percentuale di nomi e verbi si equivale (con valori tuttavia leggermente diversi, caratterizzati da uno scarto di poco più di un punto percentuale probabilmente dovuto ai diversi lessici di riferimento). Come evidenziato in Cresti (2005) in relazione all'output di uno strumento di annotazione diverso da quello usato nel presente studio ma anch'esso sviluppato in relazione alla lingua scritta, il dato relativo alla percentuale di nomi appare sovrastimato, per una naturale tendenza dello strumento di annotazione a interpretare come nomi parole sconosciute, ove ciò sia compatibile con il contesto di occorrenza. Le problematiche poste dall'annotazione del parlato sono note e non è questa la sede per soffermarsi: vale qui solo la pena menzionare che uno strumento di annotazione morfo-sintattica di trascrizioni di parlato deve sapersi confrontare con la presenza di a) forme linguistiche non standard (es. parole straniere e dialettali, produzioni onomatopeiche, neo-formazioni occasionali, parole non comprensibili), così come di b) elementi para-linguistici (es. false partenze corrispondenti a frammenti di parole, pause, "elementi riempitori" privi di contenuto semantico) ed extra-linguistici (es. tosse, riso). Utilizzando uno strumento di annotazione addestrato su corpora di lingua scritta per il parlato è inevitabile che i risultati siano falsati dalla presenza di tali elementi, che tendenzialmente vengono interpretati come nomi.

Per ripulire il risultato ottenuto dal "rumore" causato da un'analisi errata di questi elementi tipici della lingua parlata, abbiamo circoscritto l'analisi della distribuzione di nomi e verbi alle sole parole riconosciute, ovvero incluse nel dizionario morfologico di riferimento usato dallo strumento di annotazione morfo-sintattica. Le colonne 4-6 della tabella (con l'etichetta prefissata da un *) riportano la stessa tipologia di dati in relazione al sottoinsieme delle sole parole riconosciute: quanto emerge conferma che la distribuzione dei nomi era stata effettivamente sovrastimata, mostrando una preponderanza dei verbi rispetto ai nomi che si presenta come particolarmente accentuata nella porzione di corpus Parlato informale. Il rapporto nomi/verbi si è ridotto conseguentemente a valori significativamente più bassi, che riflettono la preponderanza dei verbi rispetto ai nomi (l'unica eccezione è rappresentata dal Parlato formale dove i due valori si equivalgono).

	Parlato	Parlato formale	Parlato informale	*Parlato	*Parlato formale	*Parlato informale
Nomi	17,91	19,18	16,31	15,80	17,54	13,56
Verbi	17,27	17,10	17,49	17,75	17,57	17,99
Rapporto nomi/verbi	1,04	1,12	0,93	0,89	1,00	0,75
Clausole x enunciato	1,29	1,48	1,12	-	-	-
Nomi x clausola	1,27	1,40	1,12	1,08	1,23	0,89

Tabella 7: Parametri relativi alla distribuzione di nomi e verbi nella lingua parlata basati sul testo annotato morfo-sintatticamente

Quanto registrato sia in relazione allo scritto e al parlato sia in rapporto ai diversi sotto-corpora appare in linea con la letteratura sulla variazione linguistica a livello diamesico, diafasico e testuale. E' interessante notare la coincidenza di quanto emerge dall'analisi automatica dei corpora con quanto riportato in Voghera (2004) in relazione alla distribuzione di nomi e verbi nel parlato e nello scritto: parlato 15,7% nomi vs 20,0% verbi; scritto 25,0% nomi vs 15,8% verbi. Sul versante della lingua scritta abbiamo visto che una maggiore frequenza di nomi è associata a testi caratterizzati da un'alta densità informativa (quali i giornali e le leggi), mentre generi testuali più vicini alla lingua parlata (quali la narrativa e composizioni di scrittura creativa) sono caratterizzati da una maggiore frequenza di verbi. Voghera (2005) registra lo stesso tipo di tendenza negli schemi di distribuzione dei nomi e dei verbi in corpora di testi informativi italiani. Anche Biber (1995) rileva una correlazione positiva tra la frequenza dei nomi e testi scritti ad alta densità informativa come articoli di giornale e testi accademici ("written information-focused texts"), così come tra la frequenza dei verbi e testi di scrittura creativa ("imaginative prose"): dall'analisi del corpus Lancaster/Oslo Bergen (LOB) di inglese scritto emerge che i nomi oscillano da un 26,9% in testi informativi al 20,0% in produzioni di scrittura creativa, mentre i verbi si caratterizzano per la tendenza opposta, rappresentando il 16,4% in testi informativi e il 21,9% nella scrittura creativa. Anche la variazione delle frequenze di nomi e di verbi registrata per la lingua parlata è in linea con tendenze già note e ampiamente dibattute negli studi sull'argomento: si veda in proposito Cresti (2005) e Voghera (2005).

Se da un lato la convergenza tra le misure emerse nell'ambito del presente studio e quelle riportate nella letteratura sulla variazione linguistica a livello diamesico, diafasico e testuale non ci sorprende, d'altro canto va tenuto presente che sono state ottenute in modo sostanzialmente diverso: le prime emergono dall'analisi automatica di vasti corpora di lingua scritta e parlata, le seconde sono state effettuate in relazione a corpora la cui annotazione è stata rivista manualmente (con l'unica eccezione costituita dalle analisi di Cresti su C-ORAL-ROM). Questa forte correlazione tra le due serie di dati ci incoraggia a procedere verso nuove misure, incentrate su aspetti della struttura linguistica scarsamente indagati fino ad oggi in quanto difficilmente attingibili mediante un'analisi manuale, in particolare su vasta scala.

Mantenendoci ancora al livello di annotazione morfo-sintattica, è possibile fare riferimento alle sottocategorie che suddividono la classi dei verbi in ausiliari, modali e principali. Facendo riferimento a questa sottoclassificazione, è possibile arrivare a discriminare all'interno della classe dei verbi quelli che svolgono il ruolo di testa verbale (dato ricostruito escludendo dalla classe dei verbi i modali e gli ausiliari). A partire da questa informazione combinata con il numero di periodi di cui si compone un testo, si può risalire alle clausole in cui un periodo o un enunciato mediamente si articola: tale dato è riportato nella quarta riga delle Tabelle 6 e 7 per lo scritto e il parlato rispettivamente. Si tratta di un dato nuovo e sicuramente interessante che fornisce maggiori dettagli in merito al modo in cui i verbi ricorrono all'interno del testo. Passando da una misura globale della frequenza dei verbi all'interno di un corpus alla misura della distribuzione delle teste verbali per periodo (o enunciato nel caso della lingua parlata), si nota che corpora che presentano approssimativamente la stessa

frequenza di occorrenza di verbi, come Rep (13,09%) e 2Par (13,52%), possono differire in modo significativo rispetto a questo parametro: Rep registra una media di 2,41 teste verbali per periodo contro l'1,64 di 2Par.

Disponendo del numero medio di clausole per periodo, è possibile raffinare la misura del rapporto nomi/verbi, calcolando la ricorrenza media di nomi per clausola; tale dato è riportato nell'ultima riga delle Tabelle 6 e 7. Per lo scritto, il massimo e il minimo sono osservati rispettivamente in relazione ai corpora Giur e RaccFant, mentre i valori rilevati nella lingua parlata presentano valori nettamente più bassi che si aggirano attorno al valore minimo rilevato per lo scritto. Per il parlato, il valore minimo riguarda il Parlato informale mentre sensibilmente più alto è il valore registrato per quello formale. Si tratta di una tendenza ben descritta in Voghera (2005:132), non suffragata tuttavia da misure precise: "i testi informativi […] tendono a convogliare l'informazione in un numero di clausole minore, in confronto ai testi narrativi sia parlati sia scritti. Questa tendenza comporta l'uso di sintagmi nominali più pesanti, e in particolare un più alto numero di nomi e nominalizzazioni per clausola". Mediante il ricorso alle tecnologie linguistico-computazionali questa generica tendenza può essere monitorata con precisione all'interno delle diverse varietà di lingua considerate.

Per quanto il parametro riguardante il numero medio di nomi per clausola rappresenti una misura più precisa di come nomi e verbi si distribuiscano all'interno di diverse varietà di lingua, è interessante notare che i due parametri "rapporto nomi/verbi" vs "nomi per clausola" rappresentano misure altamente correlate, con un alto indice di correlazione (r = 0,99).

	Rep	Narr	RaccFant	Suss	2Par	Giur	Media
Dip x TV	2,07	1,92	1,77	1,87	2,09	1,79	1,92
Profondità media dei livelli di incassamento in strutture nominali complesse	1,45	1,36	1,31	1,27	1,31	1,84	1,42

Tabella 8: Parametri relativi alla distribuzione intraclausale dei nomi nella lingua scritta basati sul testo annotato a dipendenze

Il numero medio di sostantivi per clausola è il massimo che si possa attingere da un testo annotato morfo-sintatticamente. La domanda che si pone a questo punto è se e in che misura questo dato sia indicativo di come nomi e verbi si rapportino gli uni agli altri all'interno della struttura sintattica. La risposta è negativa: l'annotazione morfo-sintattica non fornisce informazione alcuna in merito alle relazioni che intercorrono all'interno di ciascuna clausola tra i nomi e la testa verbale. Tale informazione può essere ricostruita solo a partire da un'annotazione a dipendenze, dove per ciascun sostantivo è possibile ricostruire la testa da cui esso è governato. In questa prospettiva, l'interrogativo da cui siamo partiti può essere riformulato come segue: per ciascuna clausola (sia essa principale o subordinata), ricostruire l'insieme dei dipendenti direttamente governati dalla sua testa verbale, di qualsiasi natura essi siano; ovvero, non solo di tipo nominale, costituiti sia da argomenti sotto-categorizzati dal verbo, sia da modificatori di varia natura (locativi, temporali, causali, etc.). Come primo passo in questa direzione, abbiamo ricostruito il numero medio di dipendenti per testa verbale

(TV); la prima riga delle Tabelle 8 (scritto) e 9 (parlato) registra tale dato per i diversi corpora considerati.

	Parlato	Parlato formale	Parlato informale
Dip x TV	1,95	2,01	1,88
Profondità media dei livelli di incassamento in strutture nominali complesse	1,37	1,44	1,24

Tabella 9: Parametri relativi alla distribuzione intraclausale dei nomi nella lingua parlata basati sul testo annotato a dipendenze

La misura dei dipendenti per testa verbale non sembra costituire un parametro discriminante per la caratterizzazione dell'opposizione scritto vs parlato in quanto i valori rilevati sono molto vicini. Appare invece un parametro utile per la caratterizzazione delle diverse classi all'interno delle due varietà diamesiche considerate. Per quanto concerne la lingua scritta, si osservano valori più alti in 2Par e Rep, mentre sul versante opposto abbiamo Giur con il valore più basso: in questo caso, il fronte dei testi altamente informativi si è spaccato con i testi giuridici che si oppongono alla prosa giornalistica di Rep e 2Par. Un andamento analogo può essere rilevato nel parlato con le categorie di Parlato formale e Parlato informale, che si collocano rispettivamente ai due estremi della scala dei possibili valori osservati per questo parametro: il primo con un valore tra i più alti (2,01), e il secondo con un valore relativamente basso (1,88).

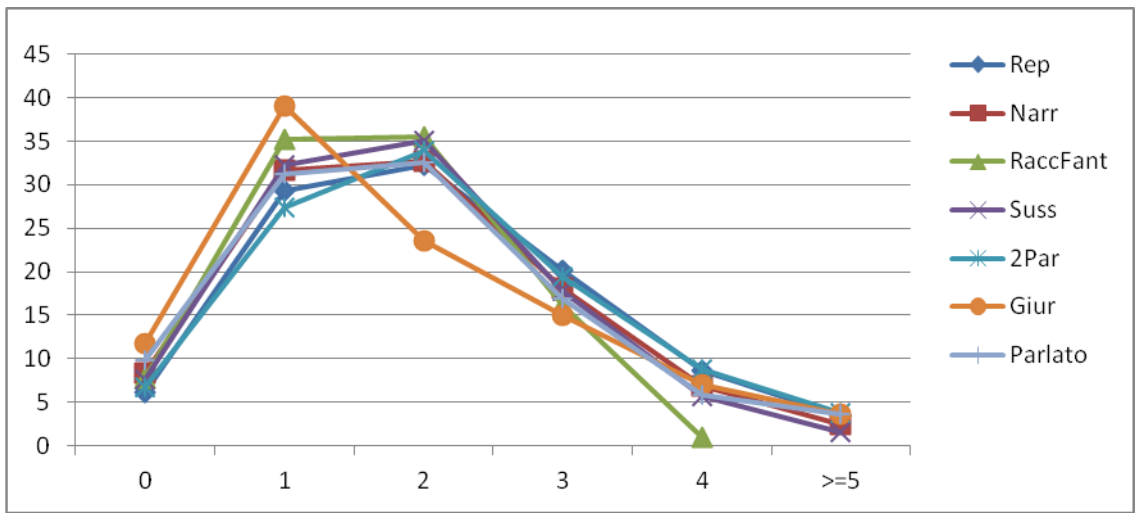

Figura 2: Distribuzione delle teste verbali per numero di dipendenti istanziati

Tale dato diventa ancor più significativo se ricostruiamo la distribuzione delle teste verbali per numero di dipendenti istanziati in modo esplicito all'interno del testo, riportata nel grafico in Figura 2. Si nota che Giur e RaccFant rappresentano i generi testuali che si allontanano maggiormente dagli altri corpora, che presentano tutti un andamento analogo. In Giur si ha un numero significativamente più alto di teste verbali

monoargomentali, e numeri più bassi di teste verbali che presentano un numero complessivo di dipendenti (argomenti o modificatori) maggiore di 1.[4] RaccFant presenta una tendenza simile ma meno marcata, caratterizzata da uno stacco minore per quanto riguarda i predicati monoargomentali e con un andamento simile per quanto riguarda gli altri.

E' interessante a questo punto verificare se e in che misura il parametro riguardante i dipendenti per testa verbale si rapporta alla misura dei nomi per clausola (vedi Tabelle 6 e 7). Confrontando i valori delle due serie di misure si osserva che il risultato differisce in modo sostanziale per lo scritto e il parlato: nel caso del primo si registra un indice di correlazione (Pearson) basso (con $r = 0,05$) mentre la correlazione è totale ($r = 1$) per quanto riguarda il parlato. Per la lingua scritta, si nota che in Giur al valore particolarmente alto di nomi per clausola (4,09 rispetto a un valore medio di 2,37) corrisponde un valore particolarmente basso di dipendenti per testa verbale (1,79 rispetto a una media di 1,92). La domanda legittima a questo punto è come possa essere spiegato questo apparente contrasto tra i due dati. Stando a Biber (1995), l'uso di nomi, derivati nominali e nominalizzazioni è tipico di testi in cui si vuole compattare molta informazione in relativamente poche parole, una caratteristica – questa – che contraddistingue testi con marcata funzione informativa. Al contrario, in testi in cui è prevalente la funzione fatica ("interactive, interpersonally-focussed on-line production") si osserva una correlazione negativa con l'uso di strutture nominali. Chiaramente questo aspetto dell'organizzazione testuale non viene catturato dalla misura dei dipendenti per testa verbale. Questo aspetto può essere utilmente monitorato mediante un altro parametro a nostro avviso importante per la caratterizzazione dei generi testuali, riguardante la profondità (o, in altri termini, "pesantezza") delle catene di dipendenza a testa nominale: se i nomi rilevati non appaiono rapportarsi direttamente alla testa verbale (ovvero, non rappresentano suoi dipendenti immediati), vanno cercati in strutture nominali complesse contraddistinte dalla presenza di modificatori (aggettivali, nominali, preposizionali).

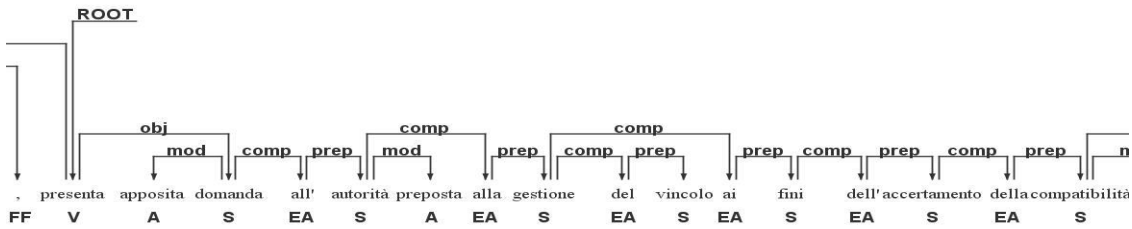

Figura 3: Esempio di struttura nominale complessa

Consideriamo il caso della modificazione di teste nominali ad opera di aggettivi e complementi preposizionali esemplificato nel frammento di albero a dipendenze in Figura 3, dove si osserva una lunga catena (interrotta nella figura) di relazioni di

[4] Le ragioni di questa peculiarità del linguaggio giuridico possono essere molteplici e necessiterebbero ulteriori indagini. Qui vale la pena menzionare due fatti. Innanzitutto, che la percentuale di soggetti espliciti rilevati in Giur è nettamente inferiore a quanto osservato in altri corpora della lingua scritta: 1,9 in Giur vs 4,5/4,7/4,7 nei corpora Rep/RaccFant/Narr. L'altro dato riguarda la frequente ricorrenza di costruzioni sintattiche caratterizzate dall'assenza di un soggetto esplicito, ad esempio le clausole participiali: tale informazione può essere inferita dalla frequenza di occorrenza di participi verbali, che rappresentano il 41,06% delle forme verbali in Giur mentre oscillano tra il 12% e il 9% negli altri corpora di lingua scritta.

dipendenza a partire dalla testa nominale *domanda* nella frase, tratta da un testo legislativo, *Il proprietario, possessore o detentore a qualsiasi titolo dell'immobile o dell'area interessati dagli interventi di cui al comma 1-ter, presenta apposita* ***domanda*** *all'autorità preposta alla gestione del vincolo ai fini dell'accertamento della compatibilità paesaggistica degli interventi medesimi.* La seconda riga delle Tabelle 8 (scritto) e 9 (parlato) riporta la profondità media delle "catene" di modificazione con testa nominale come quella esemplificata in Figura 3, inclusive di aggettivi e/o complementi preposizionali: come si può notare, si va da una profondità minima di 1,27 che si osserva in Suss a una massima di 1,84 che si ritrova, appunto, in Giur. Questo dato spiega l'apparente contraddizione dei dati rilevata in precedenza, tra il limitato numero di dipendenti per testa verbale e l'alta frequenza di nomi in testi giuridici. Anche in questo caso il parlato presenta un andamento analogo, tuttavia con valori più bassi rispetto allo scritto. Ciò appare ancor più evidente nel grafico in Figura 4 che riporta la distribuzione delle strutture nominali complesse per profondità della catena di modificazione: Giur, immediatamente seguito da Rep, presentano catene di modificazione più profonde rispetto all'altro corpus di testi informativi, ovvero 2Par, e ai testi di scrittura creativa e al parlato, caratterizzati tutti da un andamento analogo.

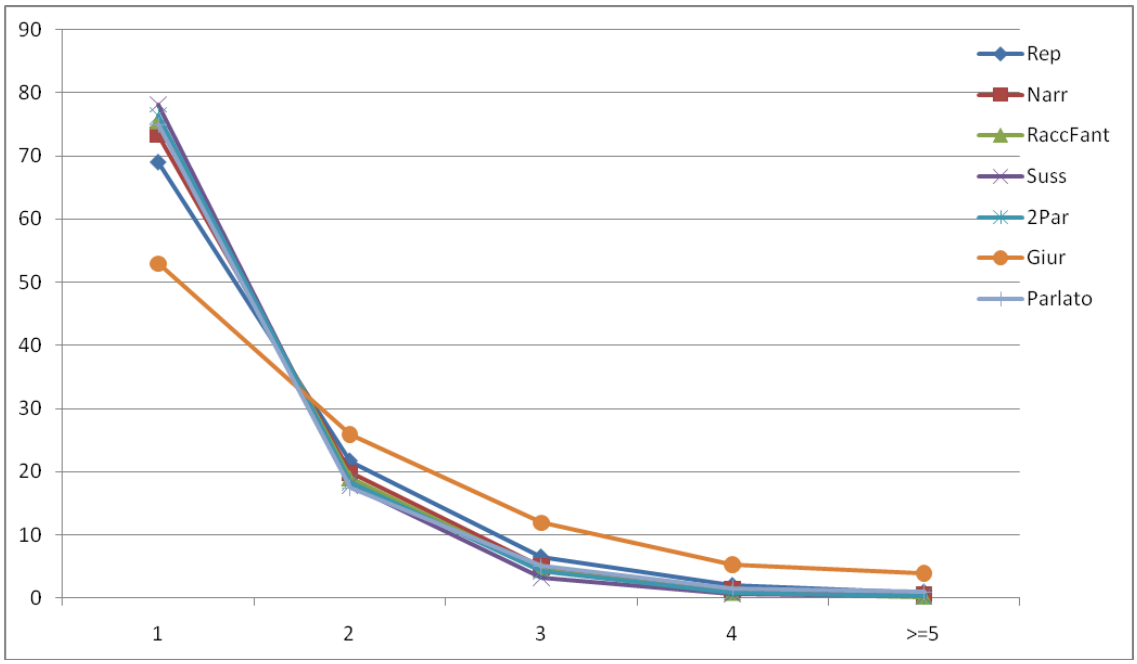

Figura 4: Distribuzione delle strutture nominali complesse per profondità

Con la possibilità di acquisire informazioni che prima rimanevano inattingibili su larga scala e in modo affidabile, anche l'ampiamente dibattuto argomento della distribuzione di nomi vs verbi in diversi generi testuali e varietà di lingua si arricchisce di ulteriore evidenza linguistica, che permette di articolare ulteriormente il quadro già noto con nuove dimensioni di analisi. In ciò sta a nostro avviso il valore aggiunto delle tecnologie linguistico-computazionali che permettono il monitoraggio di nuove dimensioni di variazione attraverso diverse varietà di lingua e testuali. Mediante questa rivisitazione di un già ampiamente esplorato caso di studio abbiamo avuto modo di

mostrare in concreto che l'orizzonte di studio può estendersi in modo significativo, andando a coprire aspetti della struttura linguistica rimasti fino ad oggi inattingibili, e dunque inesplorati.

6 QUALE RUOLO PER LE TECNOLOGIE LINGUISTICO-COMPUTAZIONALI NEL MONITORAGGIO DELLA LINGUA ITALIANA?

In questo contributo si è cercato di fornire una risposta, primariamente di carattere metodologico, all'interrogativo da cui siamo partiti, ovvero quale possa essere il ruolo delle tecnologie linguistico-computazionali nel monitoraggio della lingua italiana nelle sue diverse varietà d'uso, principalmente diamesiche, diafasiche e testuali. La risposta è positiva. Mediante il ricorso a tali tecnologie è oggi possibile monitorare in modo affidabile un ampio spettro di parametri, che spaziano tra i diversi livelli di descrizione linguistica, in relazione a corpora testuali di sempre più vaste dimensioni. Ciò può rappresentare un importante avanzamento nello studio della variazione linguistica della lingua italiana. L'approccio proposto si colloca all'interno del filone di studi di "register variation" à la Biber, con un importante elemento di novità, ovvero l'inclusione, tra i parametri di monitoraggio, di aspetti della struttura sintattica rimasti fino ad oggi inesplorati in quanto inattingibili su larga scala e in modo manuale. Il potenziale impatto di questa nuova tipologia di parametri è stato dimostrato attraverso la rivisitazione di un noto caso di studio, la distribuzione di nomi vs verbi, che essendosi arricchita di ulteriore evidenza linguistica ha permesso di articolare ulteriormente il quadro conosciuto con nuove dimensioni di analisi e variazione.

In una prospettiva più ampia, duplice risulta essere la valenza dei risultati raggiunti: sul versante teorico, la varietà delle dimensioni di analisi che possono essere simultaneamente considerate si è ampliata in modo significativo, rendendo così possibile il monitoraggio di aspetti della struttura linguistica fino ad oggi inesplorati; su un versante più di carattere applicativo, i parametri di monitoraggio identificati, o sottoinsiemi di essi, possono essere utilmente sfruttati per monitorare la competenza linguistica di apprendenti l'italiano come L1 o L2 (Dell'Orletta e Montemagni, 2012; Dell'Orletta *et al.*, 2011b), oppure per la definizione di un indice di leggibilità "avanzato" basato su parametri riguardanti l'uso della lingua in tutte le sue componenti (Dell'Orletta *et al.*, 2011a).

7 RINGRAZIAMENTI

Il presente studio è stato condotto all'interno dell'ItaliaNLP Lab dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR). Si ringraziano i membri del gruppo che hanno contribuito alla definizione della metodologia di monitoraggio proposta e hanno estratto i dati alla base del presente studio dalla selezione di corpora considerati, in particolare Felice Dell'Orletta e Giulia Venturi.

8 RIFERIMENTI BIBLIOGRAFICI

ATTARDI Giuseppe / DELL'ORLETTA Felice / SIMI Maria / TURIAN Joseph (2009) *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*. In: Evalita 2009.

BERTINETTO Pier Marco, BURANI Cristina, LAUDANNA Alessandro, MARCONI Lucia, RATTI Daniela, ROLANDO Claudia, THORNTON Anna Maria (2005) *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*, http://linguistica.sns.it/CoLFIS/CoLFIS_home.htm

BIBER, D. (1988) *Variation across speech and writing*. Cambridge & New York, Cambridge University Press.

BIBER, D. (1995) *Dimensions of register variation: A cross-linguistic comparison*. Cambridge & New York, Cambridge University Press.

BIBER D. / DAVIES M. / JONES K.J. / TRACY-VENTURA N. (2006) *Spoken and written register variation in Spanish: A multi-dimensional analysis*, "Corpora", 1, pp.7-38.

BOSCO C. / LOMBARDO V. / VASSALLO D. / LESMO L. (2000) *Building a treebank for italian: a data-driven annotation schema*, in Proceedings di LREC-2000, 2nd International Conference on Language Resources and Evaluation, Athens, pp. 99-105.

COLLINS-THOMPSON Kevyn / CALLAN Jamie (2005) *Predicting reading difficulty with statistical language models*, «Journal of the American Society for Information Science and Technology», Vol. 56, No. 13, 2005, pp.1448-1462.

CRESTI Emanuela / MONEGLIA Massimo (a cura di) (2005) *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam, John Benjamins.

CRESTI Emanuela (2005) *La testualità parlata: alcuni dati dal corpus italiano di C-ORAL-ROM nella prospettiva del parlato romanzo*. In: J. Korzen (a cura di), *Atti del VIII Convegno internazionale SILFI*, Copenaghen, giugno 2004, Copenaghen Studies, Copenaghen, pp. 163-176.

D'AGOSTINO Emilio (1998) *Il lessico di frequenza dell'italiano parlato e la didattica dell'italiano*, "Quaderns d'Italià", 3, 1998, pp. 9-28.

DELL'ORLETTA Felice (2009) *Ensemble system for Part-of-Speech tagging*. In: Evalita 2009.

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

DELL'ORLETTA Felice / MONTEMAGNI Simonetta / VENTURI Giulia (2011a) *READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification*. In: *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, pp. 73-83, Edinburgh, 30 luglio 2011.

DELL'ORLETTA Felice / MONTEMAGNI Simonetta / VECCHI Eva Maria / VENTURI Giulia (2011b) *Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria*. In: G.C. Bruno, I. Caruso, M. Sanna, I. Vellecco (a cura di), *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, Milano, McGraw-Hill, pp. 319-336.

DELMONTE R. / BRISTOT A. / TONELLI S. (2007) *VIT – Venice Italian Treebank: Syntactic and Quantitative Features*", in K. De Smedt, J. Hajič, S. Kübler (a cura di), *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*. NEALT Proceedings Series, Vol. 1, 43-54.

EVALITA (2009) *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, 12th December 2009, Reggio Emilia, Italy, 2009, http://evalita.fbk.eu/proceedings.html.

FRAZIER Lyn (1985) *Syntactic complexity*. In: D.R. Dowty, L. Karttunen e A.M. Zwicky (a cura di), *Natural Language Parsing*, Cambridge University Press, Cambridge, UK.

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

HEILMAN Michael / COLLINS-THOMPSON Kevyn / CALLAN Jamie / ESKENAZI Maxine (2007) *Combining lexical and grammatical features to improve readability measures for first and second language texts*. In: Proceedings of NAACL HLT-2007, pp.460-467.

LENCI Alessandro / MONTEMAGNI Simonetta / PIRRELLI Vito (2009) *Annotazione sintattica di corpora: aspetti metodologici*. In: C. Andorno, S. Rastelli (a cura di), *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, Perugia, Guerra Edizioni, pp. 25-46.

LIF (1972) - BORTOLINI U. / TAGLIAVINI C. / ZAMPOLLI A. *Lessico di frequenza della lingua italiana contemporanea*, Milano, Garzanti-IBM.

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

LIP (1993) - DE MAURO T. / MANCINI F. / VEDOVELLI M. / VOGHERA M. *Lessico di frequenza dell'italiano parlato*, Milano, Etas.

LU Xiaofei (2007) *Automatic measurement of syntactic complexity in child language acquisition*, «International Journal of Corpus Linguistics», 14(1), 2007, pp. 3-28.

MARCONI Lucia / OTT Michela / PESENTI Elia / RATTI Daniela / TAVELLA Mauro (1994) *Lessico Elementare*, Zanichelli, Bologna.

MARINELLI Rita / BIAGINI Lisa / BINDI Remo / GOGGI Sara / MONACHINI Monica / ORSOLINI Paola / PICCHI Eugenio / ROSSI Sergio / CALZOLARI Nicoletta / ZAMPOLLI Antonio (2003) *The Italian PAROLE corpus: an overview*. In: Zampolli Antonio *et al.* (a cura di), *Computational Linguistics in Pisa*, Special Issue, XVI-XVII, Pisa-Roma, IEPI. Tomo I, pp. 401–421.

MARINELLI Rita / BINDI Remo / MARCHI Simone / SANTARCANGELO Enrica Laura / CAVALLARO Francesca Irene / CASTELLANI Eleonora / CARLI Giancarlo (2008) *Suscettibilità ipnotica e linguaggio*, in *Atti del XLII Congresso Internazionale di Studi della Società di Linguistica Italiana* (Pisa, 25-27 Settembre 2008).

MILLER Jim / WEINERT Regina (1998) *Spontaneous spoken language. Syntax and discourse*. Oxford, Clarendon Press.

MONEGLIA Massimo (2004) *L'italiano come risorsa romanza nel corpus multilingue C-ORAL-ROM*, in Federico Albano Leoni, Francesco Cutugno, Massimo Pettorino, Renata Savy (a cura di), *Atti del convegno "Il parlato Italiano"*, Napoli, 13-15 Febbraio 2003, D'Auria Editore, Napoli.

MONTEMAGNI S. / BARSOTTI F. / BATTISTA M. / CALZOLARI N. / CORAZZARI O. / LENCI A. / ZAMPOLLI A. / FANCIULLI F. / MASSETANI M. / RAFFAELLI R. / BASILI R. / PAZIENZA M.T. / SARACINO D. / ZANZOTTO F. / MANA N. / PIANESI F. / DELMONTE R. (2003) *Building the Italian Syntactic-Semantic Treebank*, in A. Abeillé, *Treebanks*: *Building and Using Parsed Corpora*, Dordrecht/Boston/London, Kluwer Academic Publisher, pp. 189- 210.

PANUNZI Alessandro, PICCHI Eugenio, MONEGLIA Massimo (2004) *Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-Oral-Rom Italian*, LREC, in Proceedings of LREC 2004 "Fourth International Conference on Language Resources and Evaluation", Lisbon (Portugal), 26-28 May 2004, Vol. II, pp. 563-568.

PETERSEN Sarah E. / OSTENDORF Mari (2009) *A machine learning approach to reading level assessment*. «Computer Speech and Language», 23, pp. 89-106.

PIEMONTESE Maria Emanuela (1996) *Capire e farsi capire. Teorie e tecniche della scrittura controllata*, Napoli, Tecnodid.

POLICARPI Gianna / ROMBI Maggi (2005) *Tendenze nella sintassi dell'italiano contemporaneo*. In: Chiari Isabella / De Mauro Tullio *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne, Roma, pp.139-156.

ROARK Brian / MITCHELL Margaret / HOLLINGSHEAD Kristy (2007) *Syntactic complexity measures for detecting mild cognitive impairment*, in Proc. ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP'07), pp.1-8.

ROMBI Maggi / POLICARPI Gianna (1985) *Mutamenti sintattici nell'italiano contemporaneo: il sistema delle congiunzioni*, in Agostiniani, L., Bellucci Maffei, P. & Paoli, M. (a cura di), *Linguistica storica e cambiamento linguistico*, Atti SLI, 23. Roma, Bulzoni, pp. 225-244.

SAGAE Kenji / LAVIE Alon / MACWHINNEY Brian (2005) *Automatic measurement of syntactic development in child language*, in Proceedings of the 43rd Annual Meeting of the ACL.

SCHWARM Sarah E. / OSTENDORF Mari (2005) *Reading level assessment using support vector machines and statistical language models*, in Proceedings of ACL-2005, pp.523-530.

SINCLAIR John (1991) *Corpus, Concordance, Collocation*, Oxford University Press, Oxford, UK.

VELI (1989) - DE MAURO T. / IBM *Vocabolario Elettronico della Lingua Italiana*, Centro di Ricerca IBM, Roma.

VENTURI Giulia (2006) *L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale*, Tesi di Laurea Specialistica, Università di Pisa, Dicembre 2006.

VOGHERA Miriam (2001) *Riflessioni su semplificazione, complessità e modalità di trasmissione: sintassi e semantica*, in Dardano Maurizio *et al.* (a cura di), *Scritto e parlato. Metodi, testi e contesti*, Atti del Colloquio Internazionale di Studi, Aracne, Roma, pp.65-78.

VOGHERA Miriam (2004) *La distribuzione delle parti del discorso nel parlato e nello scritto*, in Van Deyck R, Sornicola R. et Kabatèk J.(a cura di), *La variabilité en langue, I. Langue parlée et langue écrite dans le présent et dans le passé, II. Les quatre variations*, Gand, Communication & Cognition (Studies in Language, 8), pp. 261-284.

VOGHERA Miriam (2005) *La misura delle categorie sintattiche*, in Chiari Isabella / De Mauro Tullio (a cura di) *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne, Roma, pp.125-138.

YNGVE Victor H.A. (1960) *A model and an hypothesis for language structure*. In: *Proceedings of the American Philosophical Society*, pp. 444-466.