LA BANCA-DATI DEI PROVVEDIMENTI DELLA SEZIONE DISCIPLINARE DEL CONSIGLIO SUPERIORE DELLA MAGISTRATURA (1990-2007)

Manuela Sassi e Alessandra Cinini (ILC-CNR)

1. Introduzione

L'Istituto di Linguistica Computazionale (ILC-CNR di Pisa) ha collaborato con l'IRSIG-CNR e con il Centro Studi e Ricerche sull'Ordinamento Giudiziario (CESROG) dell'Università di Bologna per lo studio e la realizzazione di una banca-dati elettronica dei provvedimenti della Sezione Disciplinare del Consiglio Superiore della Magistratura. La collaborazione si è sviluppata sul tema seguente: "Individuazione ed estrazione di informazioni da banche dati giuridiche tramite lo studio e la creazione di regole testuali con parametri di tipo linguistico. Creazione e sperimentazione di protipi per l'analisi della giurisprudenza"; ciascun gruppo di lavoro ha contribuito all'obiettivo apportando l'esperienza della propria area disciplinare, al fine di costruire un percorso di ricerca interdisciplinare, dove si fondono conoscenze giuridiche, informatiche e linguistiche. In questo articolo verranno affrontate più in dettaglio le tematiche linguistico-informatiche, che comprendono già, ad un primo livello di analisi, le conoscenze giuridiche apportate dai collaboratori dell'IRSIG e del CESROG. Queste ultime inoltre rappresentano lo strumento essenziale al momento di svolgere le più approfondite analisi sui contenuti, una volta che la banca-dati sia stabile e rappresentativa dell'attività della Sezione Disciplinare del CSM, in anni molto significativi per la storia della Magistratura e della Repubblica Italiana.

L'ILC ha le competenze e la disponibilità di strumenti efficaci, già ampiamente sperimentati nella ricerca linguistica (Picchi, 1999; Sassi, 2003); nell'ambito giuridico possono avere una applicazione proficua che, nonostante sia stata intravista molto precocemente negli anni passati, in realtà non ha avuto un grande sviluppo. Con questo progetto, a partire dal primo prototipo di banca-dati realizzato con i fondi FIRB, e continuato negli anni successivi, si vuole dimostrare la capacità e la versatilità di tali strumenti linguistico-computazionali, già sviluppati e testati nel mondo delle discipline letterarie (Ceccotti e Sassi, 2003), per lo studio del linguaggio giuridico nelle sue varie specializzazioni, e soprattutto per l'estrazione di informazioni linguistico-concettuali da grandi quantità di dati testuali, tali che favoriscano e consentano operazioni di sintesi di più ampio respiro.

2. Prototipo della banca dati CSM

Il primo obiettivo del progetto è stata la realizzazione del prototipo della banca-dati elettronica dei provvedimenti della sezione disciplinare del Consiglio Superiore della Magistratura.
L'IRSIG e il CESROG ci hanno inviato i testi dei provvedimenti emessi dalla Sezione Disciplinare, che erano stati messi a disposizione dal CSM.
I testi del primo invio, sotto forma di archivio storico dei files in vari formati, consistevano di oltre 1600 documenti. Su tali materiali si è operato in modo da renderli uniformi e quindi trattabili con uno stesso software. Questo trattamento è stato poi ripetuto ai testi dei due aggiornamenti successivi, che erano tutti in formato Word, e che hanno richiesto quindi meno elaborazioni per ridurli ad un unico formato testo.

2.1.Fasi di lavoro

Le fasi lavoro effettuate sono le seguenti:

(a) studio della tipologia dei testi e trascodifica dal formato di partenza al formato-testo (a partire da Wordstar, WordPerfect, RTF, Word, ecc.)
(b) codifica degli elementi individuati come prioritari, tramite uno studio previo della struttura del testo del provvedimento-tipo;
(c) applicazione ai testi, già codificati, dei moduli informatici che creano il database e i relativi indici;
(d) verifica dell'implementazione effettuata attraverso consultazioni mirate e volte ad individuare la migliore codifica dei fenomeni rilevanti come, ad esempio, la struttura, le date, i nomi propri del

collegio, dell'incolpato, dei testimoni, e di quant'altro ritenuto necessario dal gruppo di lavoro;

(e) sviluppo di software ad hoc finalizzato al riconoscimento o alla normalizzazione dei dati (es.: riduzione delle date di emissione ad un medesimo formato, estrazione dell'anno di inizio procedimento o di deposito, riconoscimento delle sezioni formali del documento, degli acronimi, delle sigle, dei numeri, dell'esito, ecc.);

(f) rigenerazione degli archivi ogni qualvolta si sia reso necessario, per modifiche nella codifica, per nuovi inserimenti e aggiornamenti del database.

(g) strategie di consultazione del database allo scopo di produrre report di content-analysis;

(h) elaborazione dei risultati delle query, sia con software ad hoc, sia con comuni prodotti di office automation, in modo da produrre la documentazione necessaria.

(i) Applicazione di metodologie di POS-TAGGING per il riconoscimento morfologico-sintattico delle parti del discorso e etichettatura di tutte le parole del database per ulteriori analisi di tipo contenutistico.

2.2.Studio della tipologia dei testi

Con i testi informatizzati delle Sentenze e delle Ordinanze della Sezione Disciplinare del Consiglio Superiore della Magistratura è stata costituita una prima banca-dati testuale utilizzando il sistema DataBaseTestuale, (DBT, Brevetto CNR, Picchi, 1997). Tale raccolta è stata successivamente integrata con vari aggiornamenti, (Cinini e Sassi, 2004) di cui il più recente copre fino al settembre dell'anno 2007.

La versione attuale comprende un insieme di 2061 provvedimenti disciplinari, di cui 184 Ordinanze e 1877 Sentenze. Inizialmente la raccolta doveva contenere i provvedimenti emessi a partire dal 1990, ma dato che nel primo invio erano compresi provvedimenti conclusi anche negli anni precedenti, questi sono stati inseriti, anche per favorire un futuro aggiornamento a ritroso nel tempo. Di contro, dopo aver inserito tutti i provvedimenti delle successive integrazioni, abbiamo effettuato un controllo con il modulo del DBT che segnala i documenti uguali, rilevando che alcuni provvedimenti erano ripetuti e quindi sono stati eliminati dalla banca-dati.

Per una descrizione più approfondita delle caratteristiche dei testi rimandiamo ai paragrafi successivi.

2.3.Creazione del database

I 2061 documenti sono stati ordinati per numero di procedimento/anno di inizio e indicizzati in modo tale che nelle risposte alle richieste venga sempre rispettato il criterio temporale. In questa fase vengono generati gli archivi in modo tale da ottimizzare sia lo spazio di memoria occupato, sia la velocità di recupero dell'informazione. Così il testo viene "predigerito" e predisposto in modo tale che, con un semplice tocco del mouse, ogni parola diventa un link, facilitando la navigazione nei documenti e nelle loro sezioni.

Sono stati prodotti più archivi-DBT, organizzati poi in un Corpus, in modo che possano essere consultati sia singolarmente che congiuntamente (in quest'ultimo modo è possibile fare una sola richiesta che produrrà risultati distinti per ciascun archivio). Per questo motivo parliamo di banca-dati, visto che l'insieme degli archivi, per omogeneità di trattamento, può essere visto come una unica fonte di dati.

Sono state elaborate un totale di oltre 4 milioni e mezzo di parole delle Sentenze e oltre duecentomila parole delle Ordinanze.

3. Caratteristiche del database

L'indicizzazione ha permesso l'applicazione di funzioni di analisi testuale agli archivi ottenuti quali:

(a) la ricerca libera nel testo di una o più parole o porzioni di esse, collegate in AND oppure in OR;
(b) la ricerca per lemma che, tramite il motore morfologico, recupera tutte le forme del lemma dato e/o dei suoi sinonimi;
(c) il calcolo delle frequenze assolute in ordine alfabetico o decrescente;
(d) le concordanze di tutte le parole secondo l'ordine di presenza nei testi, oppure secondo l'ordine alfabetico delle forme che nella frase seguono/precedono la parola oggetto di ricerca;
(e) l'index locorum, arricchito da riferimenti puntuali per tutte le parole del testo;
(f) il calcolo dell'indice di "mutual information" ossia la formula statistica che applicata ad un testo genera un elenco di coppie di parole, dette co-occorrenze, che sono collegate da uno stretto legame di correlazione nei contesti recuperati;
(g) il calcolo dell'indice di leggibilità, basato sulle formule statistiche più accreditate per l'italiano.

3.1.Codifiche e marcature ad hoc

Nell'esplorazione dell'archivio tramite le funzioni elencate nel precedente paragrafo, si sono evidenziati casi in cui era possibile intervenire con software da sviluppare ad hoc. Con questa ulteriore fase sono stati isolati e contrassegnati alcuni fenomeni testuali per migliorare la navigazione nei testi e per il recupero di informazioni utili a analisi di tipo concettuale (content-analysis).

3.1.1.I riferimenti

Ciascun provvedimento costituisce l'unità logica di testo del database e viene etichettato tramite una intestazione. Nella prima versione del database, il *riferimento logico*, ossia l'intestazione, riproduceva il nome del file originale. Ad esempio: SEN1.90 => SEN1_90 oppure SEN73.01.DOC => SEN73_01. Per procedere al loro giusto ordinamento, sono stati inseriti gli zeri non significativi e quindi i riferimenti precedenti sono diventati SEN001_90 e SEN073_01. Inoltre si è ritenuto opportuno definire una struttura standard per tale etichetta: i primi tre caratteri del riferimento logico specificano il tipo di provvedimento, seguono il numero di procedimento, il simbolo "_" e l'anno di inizio (nel caso che un provvedimento riunisca più procedimenti, questi sono separati dal segno "+"), gli eventuali nomi o termini descrittivi sono delimitati dal segno "-". Seguono alcuni riferimenti che esemplificano la casistica descritta:

SEN074_03+123_04
SEN005-P_96
ORD010_06
ORD011-S_95

In seguito il riferimento è stato arricchito da ulteriori informazioni che verranno descritti nei paragrafi successivi.

3.1.2.Le date

Studiando le peculiarità dei documenti ci siamo resi conto delle varianti nell'indicazione della data di emissione. Dopo un'attenta analisi di tali varianti, è stato studiato un programma per la marcatura automatica della data di emissione del provvedimento disciplinare, per poi procedere alla sua normalizzazione e renderla così utilizzabile come parametro nella fase di consultazione. Si esemplificano alcuni casi di diverse notazioni:

Roma, 19.10.1990
Roma, 28 settembre 1990

Roma, 16/11/1990
Roma, 26/10/90
Roma, 1° marzo 1991
Roma, 1°.7.1991

Il riconoscimento della data ha permesso successivamente di arricchire il riferimento logico con l'inserimento dell'anno di emissione del provvedimento disciplinare. L'idea di introdurre questa ulteriore codifica è apparsa molto utile nella fase di consultazione di prova degli archivi; a seguito di questa integrazione il reperimento dei procedimenti conclusi in un determinato intervallo risulta più semplice e veloce.
Con tale marcatura il riferimento risulta così strutturato:

Tabella 1 - Struttura del riferimento

Sen/Ord	*N.Proc.*	*Anno Inizio*	*(Anno Conclusione)*
SEN	045	_92	(95)
ORD	002	_93	(95)

Di seguito si propone l'analisi ottenuta con l'estrazione dei riferimenti dal risultato della richiesta di tutte le date di emissione dell'archivio delle Sentenze, ottenuta con i seguenti passi:

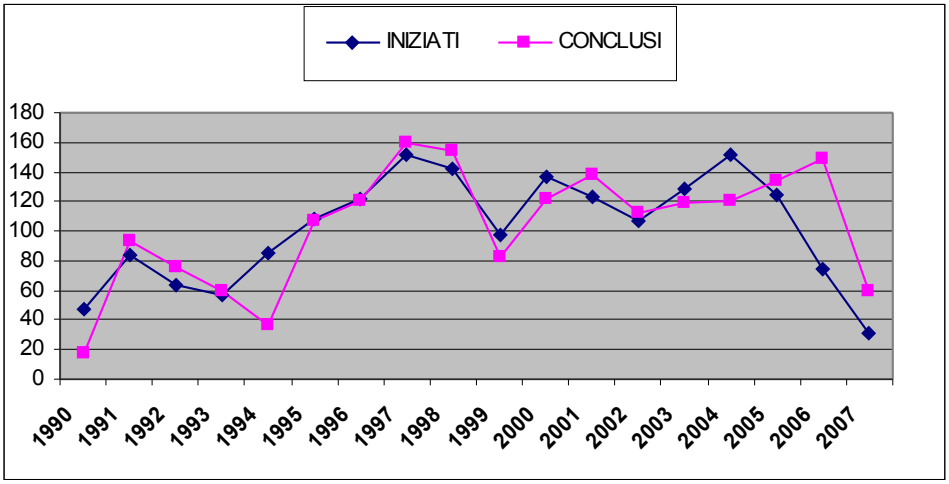
- Il risultato della query è stato salvato su file con la funzione SalvaFinestra (il programma di consultazione consente il salvataggio delle finestre dei risultati accodandole in uno stesso file).
- Tale file è stato composto in forma di tabella di tutte le date di emissione dei procedimenti. Infatti si è rilevato che in 1869 Sentenze è presente la data di conclusione, mentre quella di deposito era mancante perché probabilmente è stata apposta manualmente nelle copie cartacee; per questo motivo il calcolo della durata media dei procedimenti è stato effettuato in base alla data di emissione della sentenza.
- I dati sono stati poi trattati con un foglio elettronico in modo da fare i calcoli con funzioni apposite e darne così anche una visione grafica.

Nella tabella 2 viene illustrato l'andamento del numero di procedimenti disciplinari iniziati e conclusi per ogni anno, dal 1983 al 2007 (quest'ultimo anno non è completo, come quelli prima del 1990).
Nella figura 1 viene rappresentato graficamente l'andamento dei procedimenti a partire dal 1990.

Tabella 2 - Elenco dei procedimenti per anno

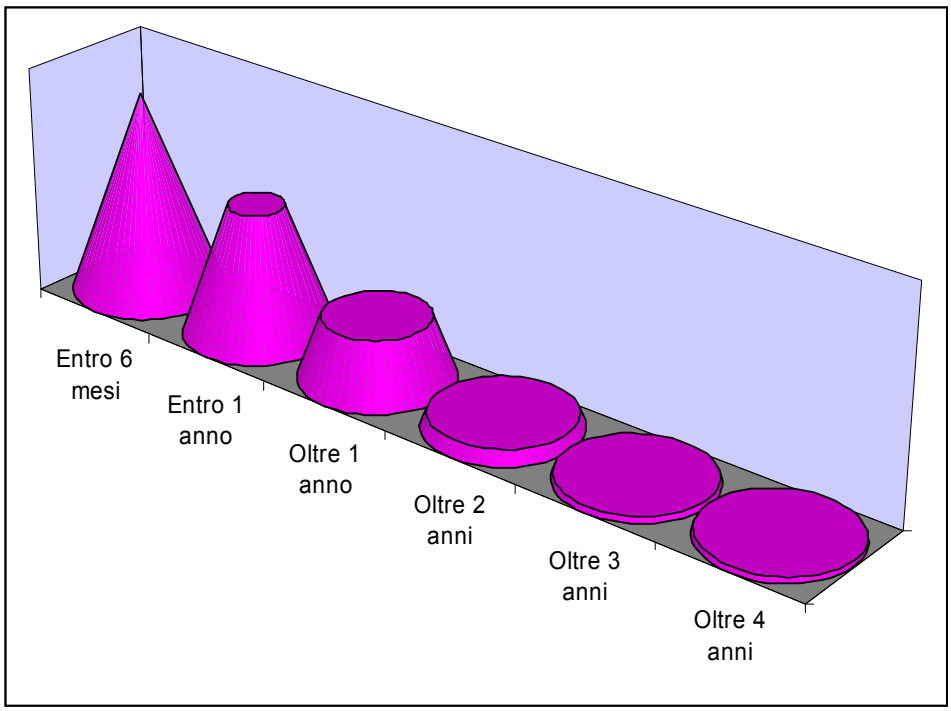
ANNO	*PROC. INIZIATI*	*PROC. CONCLUSI*
1983	1	0
1985	1	0
1986	2	0
1987	11	1
1988	18	6
1989	11	1
1990	46	17
1991	84	93
1992	64	76
1993	57	59
1994	85	36
1995	107	107
1996	123	120
1997	151	160
1998	142	154
1999	98	83
2000	137	122
2001	123	138
2002	107	113
2003	128	119
2004	151	121
2005	125	134
2006	74	149
2007	31	60
Totale	1877	1869

Figura 1 - Procedimenti iniziati/conclusi per anno



Basandosi sempre sul risultato salvato nella ricerca precedente e sul contenuto del riferimento, è stato svolto uno studio sulla durata dei procedimenti disciplinari delle Sentenze. La durata è stato ottenuta calcolando in semestri l'intervallo di tempo tra la data di inizio e quella di conclusione; la data di inizio è stata ricostruita tramite una formula basata sul numero di procedimento disciplinare, presente nel riferimento logico. La durata media di un procedimento disciplinare è risultata di poco superiore a 8 mesi. Sono state individuate sei classi per la valutazione della durata dei procedimenti e la figura 2 illustra la quantità di procedimenti per ciascuna classe di durata.

Figura 2 - Durata dei procedimenti emessi

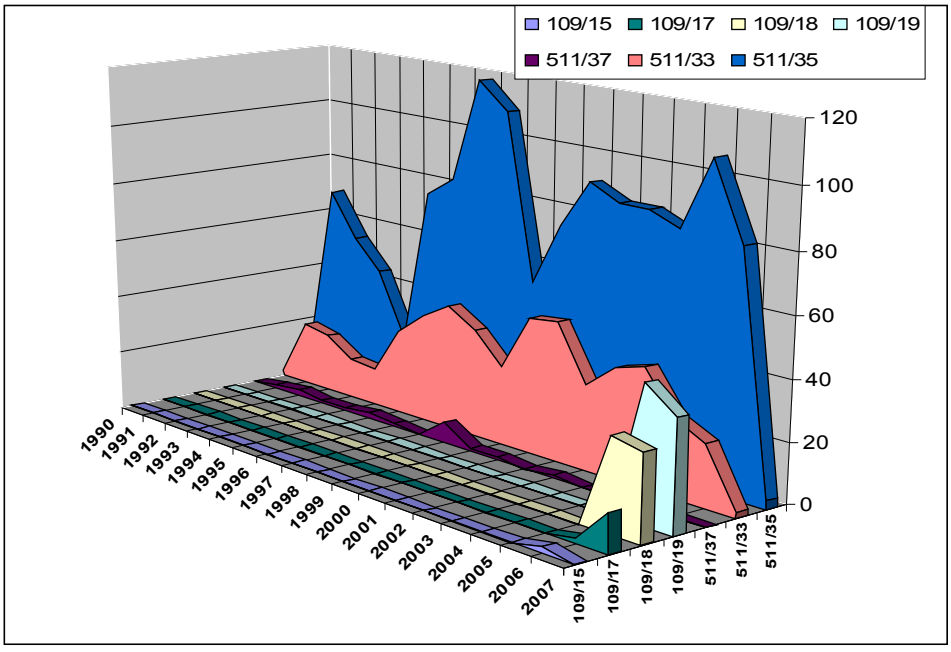


3.1.3.I numeri

Un'altra operazione ritenuta utile per la consultazione della banca-dati è stata individuata nella marcatura dei numeri presenti nel testo.
Con questa codifica si possono, ad esempio, recuperare tutte le sentenze che citano un determinato articolo di legge, oppure che fanno riferimento a determinate quantità: monetarie, di misura, di tempo, ecc.

Per illustrarne le possibilità di utilizzo, si propone un grafico che quantifica il numero di sentenze nel cui PQM si citano gli articoli del R.D.L. n.511 e quelli del d.l. n.109 del 2006. Nella figura 3 sono state rappresentate le citazioni degli articoli che compaiono all'interno del PQM nella banca-dati delle Sentenze, suddivise per anno e per articolo, tralasciando quelli con frequenza minore di 5.

Figura 3 Quantità di citazioni nel PQM degli artt. del 511 e del 109 per anno



3.1.4.Le sigle o acronimi

Un'ulteriore trasformazione effettuata sugli archivi testuali è stata la codifica delle abbreviazioni, molte delle quali sono state utilizzate come forme di citazioni di atti normativi, e degli acronimi, che si riferiscono generalmente ad enti, istituti, società, ecc., coinvolti a vario titolo nel provvedimento.

L'individuazione degli acronimi e delle abbreviazioni nei documenti è stata realizzata con l'ausilio di un modulo del DBT denominato "CercaFenomeni", che utilizza elementi come le lettere maiuscole e/o la presenza di punti non seguiti da maiuscola; la codifica è stata effettuata

aggiungendo una marcatura iniziale e finale alla forma presente nel documento.
L'elenco di tutte le sigle codificate può essere rapidamente generato in fase di consultazione usando il canale di ricerca apposito (esiste la voce specifica "Ricerca Sigle").
Nell'esame del risultato della ricerca delle sigle si rileva una difformità nelle abbreviazioni usate e tale lista (di cui mostriamo il frammento relativo a *decreto*) potrebbe essere uno strumento utile per uniformare le abbreviazioni e gli acronimi presenti nei testi secondo degli standard prestabiliti.

d. lgs.	d.d. m.m.	decr. lg.vo	decr.
decr. leg.	decr. leg.vo	decr. leg.vo	dlgs.
dd.ll.	d.l.	d.l.g.	d.l.s
d.l.v.	d.l.vo	d.leg.vo	d.lg.
d.lg.vo	d.lgs	d.lgs.	d.lgs.vo
d.lgt.	d.lvo	d.lvo.	d.m.
d.p.	d.p.c.m.	d.p.r.	d.r.lgs.

3.2.La struttura delle sentenze

Tutti i provvedimenti sono impostati secondo una struttura fissa; allo scopo di recuperare tale struttura a livello informativo, si è proceduto alla individuazione dei segnali che indicano i livelli o sezioni di tale struttura.
In alcuni casi sono chiaramente indicati nel testo: ad esempio la sigla P.Q.M. (o p.q.m. dato che, a livello di etichetta, il DBT non differenzia maiuscole e minuscole) individua il dispositivo, quella parte della sentenza che contiene la decisione del caso; per identificare tale parte è stato usato il marcatore \PQM\.
Nella sezione dedicata all'esposizione dei fatti e dei motivi, vi sono diversi modi di organizzazione del testo: è stato scelto di lasciare divise le due sezioni, qualora ciò sia indicato chiaramente nel testo, oppure di non distinguerle, nei casi in cui non vi siano segnali inequivocabili.
In mancanza di segnali non viene introdotto nessun marcatore, mentre in presenza dei seguenti segnalatori, viene inserito il marcatore corrispondente, come risulta dalla tabella 3, dove vengono elencate le sezioni codificate:

Tabella 3 - Corrispondenza fra sezione del testo e marcatore

\FATTO	Svolgimento del processo (e simili)
\DIRITTO	Motivi della decisione (e simili)
\FATTO_DIRITTO	Precedenti sezioni indistinte
\SENTENZA	Sentenza
\ORDINANZA	Ordinanza
\INCOLP.	Incolpato/a/e/i

In fase di ricerca si potrà quindi utilizzare un ulteriore canale di accesso all'informazione secondo il marcatore di struttura: infatti tutto il testo tra un marcatore e l'altro verrà attribuito al marcatore corrispondente; ciò rende quindi possibile delimitare la ricerca solamente a quei contesti che si trovino all'interno della sezione indicata tramite la funzione di "Restrizione", disponibile nel menu del programma di consultazione.
In occasione dell'ultima revisione della banca-dati, con la quale è stato fatto anche un significativo aggiornamento dei provvedimenti emessi fino a coprire parzialmente l'anno 2007, è stato approfondito lo studio delle formule che "aprono" le sezioni FATTO e DIRITTO.
Nella Tabella 4 si elencano tali formule e il corrispondente marcatore applicato:

Tabella 4 - Elenco delle formule di apertura delle sezioni FATTO e DIRITTO

\FATTO_DIRITTO	fatto e diritto
	fatto e svolgimento del processo
	in fatto e diritto
	in fatto e in diritto
	in fatto ed in diritto
	osserva
	ritenuto in fatto e considerato in diritto
	ritenuto in fatto e diritto
	ritenuto in fatto e in diritto
	svolgimento del procedimento
	svolgimento del procedimento e motivazione della decisione
	svolgimento del procedimento e valutazioni in fatto ed in
	svolgimento del processo e motivazione della decisione
	svolgimento del processo e motivi della decisione
\FATTO	delimitazione dei fatti oggetto del giudizio di rinvio
	fatto
	in fatto
	premesso
	premesso che
	premesso in fatto
	rileva in fatto

	ritenuto in fatto
	svolgimento del fatto
	svolgimento del procedimento
	svolgimento del processo
\DIRITTO	diritto
	diritto e motivi della decisione
	in diritto
	motivazione
	motivazione della decisione
	motivazione della sentenza
	motivi
	motivi della decisione
	osserva
	osserva in diritto
	rilevato
	ritenuto in diritto
	si osserva

4. Esempi di consultazione della banca-dati

Con l'inserimento delle varie tipologie di codici nei testi di input, come precedentemente descritto, i dati vengono trasformati in informazioni che possono essere selezionate ed estratte tramite le funzioni di ricerca del DBT.

I contesti, raccolti e salvati su file, vengono successivamente elaborati e trattati come dati etichettati e possono essere sintetizzati con programmi di calcolo, che ne permettono anche delle elaborazioni grafiche, oppure possono servire ad ulteriori approfondimenti della codifica.

Ad esempio, al fine di individuare la tipologia delle sanzioni inflitte nell'anno 1995 e la loro frequenza di applicazione in quell'anno, sono stati richiesti tutti i contesti in cui è stato usato il verbo *infliggere* o *deliberare* (in tutte le forme coniugate tramite la funzione di ricerca per lemma) che fossero presenti nella sezione marcata con *PQM*; a questa ricerca è stata aggiunta la condizione che il riferimento logico contenesse la stringa alfanumerica *(95)* tra parentesi, in quanto indica l'anno di conclusione. I contesti sono stati ordinati secondo le parole che compongono la formula di rito e raggruppati in base all'esito. I risultati sono stati organizzati nella tabella 5.

Tabella 5 - Raggruppamento delle sentenze del '95 per esito

AMMONIMENTO	*CENSURA*	*PERDITA DELL'ANZIANITA'*	*RIMOZIONE*	*DISPENSA DALL'UFFICIO*
SEN005_94(95)	SEN047_94(95)	SEN034_94(95)	SEN096_94(95)	SEN022_94(95)
SEN012_93(95)	SEN019_95(95)	SEN031_93(95)		
SEN015_95(95)	SEN021_95(95)	SEN101_94(95)		
SEN024_95(95)	SEN007_94(95)			
SEN028_95(95)	SEN071_93(95)			
SEN040_94(95)	SEN060_93(95)			
SEN042_95(95)	SEN035_94(95)			
SEN045_92(95)	SEN046_94(95)			
SEN060_94(95)	SEN045_94(95)			
SEN061_94(95)	SEN070_94(95)			
SEN062_93(95)	SEN016_93(95)			
SEN064_94(95)	SEN089_94(95)			
SEN078_94(95)	SEN016-4_93(95)			
SEN090_94(95)				
SEN006_94(95)				
SEN025_95(95)				
SEN055_94(95)				
SEN067_94(95)				
SEN073_94(95)				
SEN085_94(95)				

A seguito di ulteriori elaborazioni dei risultati di query simili alla precedente, è stato individuato l'esito di ogni sentenza. Ovviamente questo lavoro è stato reso possibile dalla continua interazione fra i collaboratori del progetto, che ad ogni passo hanno contribuito con le loro specifiche competenze. Ad esempio, mantenendo impostata la restrizione sulla sezione che limita la ricerca al dispositivo, sono state richieste altre parole-chiave, come *assoluzione* e *assolvere*, o famiglie di parole, come *sanzione & ammonimento*, *non & rinvio & dibattimento*, *non & doversi & procedere*, ecc. Le finestre relative ai contesti recuperati sono state salvate e organizzate in ordine alfabetico in modo da raggruppare le diciture omogenee. A partire da questo elenco di formule gli specialisti hanno suggerito la codifica standard di ogni esito, che è stata poi fornita in input ad una procedura automatica, che ha aggiunto alla fine di ogni riferimento logico l'esito o gli esiti recuperati dai contesti corrispondenti.
Nella versione finale della banca-dati quindi abbiamo ottenuto un riferimento logico composto come riportato in Tabella 6.

Tabella 6 - Struttura finale del riferimento logico

Sen	*N.Proc.*	*Anno Inizio*	*(Anno Conclusione)*	*Esito*
SEN	045	_92	(95)	**AMM**
SEN	094	_96	(96)	**ASS/CEN**
SEN	034	_90	(91)	**AMM/ASS**

Come si può notare dalla precedente Tabella 6, vi sono provvedimenti che hanno più esiti, in relazione ai capi di incolpazione; gli eventuali doppi esiti utilizzano la barra come separatore, e ciascun esito è composto da un codice di tre caratteri qui elencati:

AMM	AMR	ASS	CEN	DDU	DES	EAD
EPM	EPO	ERM	EST	FLD	FRD	INR
NDP	NLG	NLP	NRD	P_A	RIM	

Tali codici possono essere usati con la funzione di Restrizione durante la fase di consultazione della banca-dati, in modo da selezionare ulteriormente i provvedimenti oggetto della ricerca.

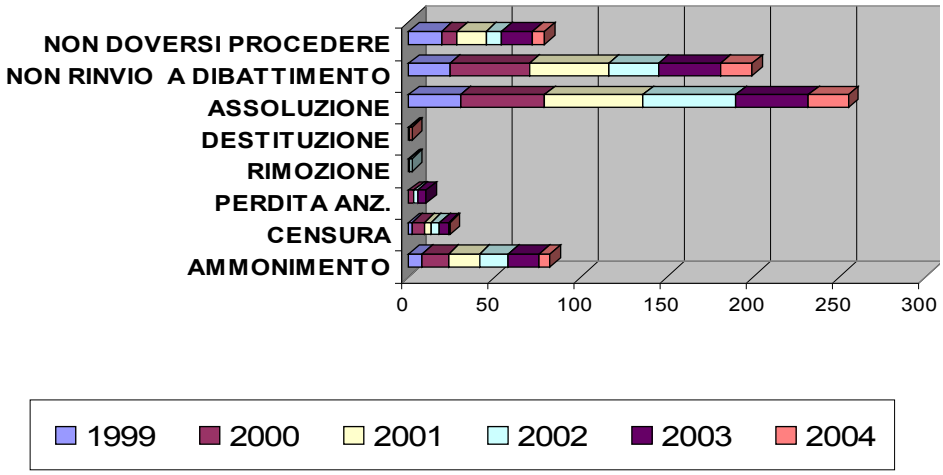
Questa elaborazione ha permesso di poter stilare una tabella generale di tutti i provvedimenti organizzati secondo l'anno di conclusione e l'esito, come si può notare nella tabella 7, dove la codifica presente nel riferimento è stata resa più esplicita. I provvedimenti con esiti multipli sono stati conteggiati in numero pari ai vari esiti.

Tabella 7- Procedimenti suddivisi per esito

	90	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**00**	**01**	**02**	**03**	**04**	**05**	**06**	**07**	
AMMON	3	17	11	11	4	20	12	19	16	8	15	18	17	18	15	19	15	7	245
CENSURA	1	15	6	7	4	13	20	17	8	2	7	4	5	5	4	8	8	3	137
PERD. ANZ	1	5	6	2		3	4	2	4		3		2	5	2	5	7	5	56
RIMOZIONE				2	1	1			1		1		1						7
DESTIT.	1		1										1		1	1	1		6
TOT CON	**6**	**37**	**24**	**22**	**9**	**37**	**36**	**38**	**29**	**10**	**26**	**22**	**26**	**28**	**22**	**33**	**31**	**15**	**451**
ASSOL.	7	26	21	20	9	34	24	59	41	31	49	57	54	42	48	63	78	19	682
NON RINV DIB	2	22	19	12	12	28	36	43	37	26	46	47	29	36	41	27	23	9	495
NON DOV PROC	2	13	15	10	8	16	30	36	49	19	8	17	9	18	15	17	26	19	327
TOT NON CON	**11**	**61**	**55**	**42**	**29**	**78**	**90**	**138**	**127**	**76**	**103**	**121**	**92**	**96**	**104**	**107**	**127**	**47**	**1504**
ALTRI	0	3	4	2	1	3	4	0	7	2	1	1	1	0	0	1	0	2	32

Allo scopo di rendere in modo grafico i dati numerici presentati nella precedente tabella, è stato estrapolato il quinquennio 1999-2004 e organizzato nella Figura 4, dove possiamo vedere le categorie degli esiti in formato descrittivo.

Figura 4 - Grafico degli esiti estratti dalla tabella 7(1999-2004)



4.1.Indici vari

Oltre alla ricerca diretta di informazioni tramite l'interfaccia di consultazione dei testi, il sistema DBT consente la creazione di numerosi indici organizzati in files di output, riutilizzabili con un qualunque programma di videoscrittura:

- gli index locorum, arricchiti da riferimenti puntuali,
- la lista di forme co-occorrenti con i relativi valori di *mutual information*,
- il calcolo delle frequenze assolute in ordine alfabetico o decrescente,
- il calcolo dell' indice di leggibilità, basato sulle formule statistiche più accreditate.

4.1.1.Index locorum

Per dare un esempio di questo tipo di risultato si estrapola dall'index locorum completo delle Ordinanze il seguente frammento, dove si elenca la frequenza e i luoghi di ogni parola. I nomi che compaiono nel riferimento sono stati mascherati (x, y e z) e l'ultimo numero riporta la

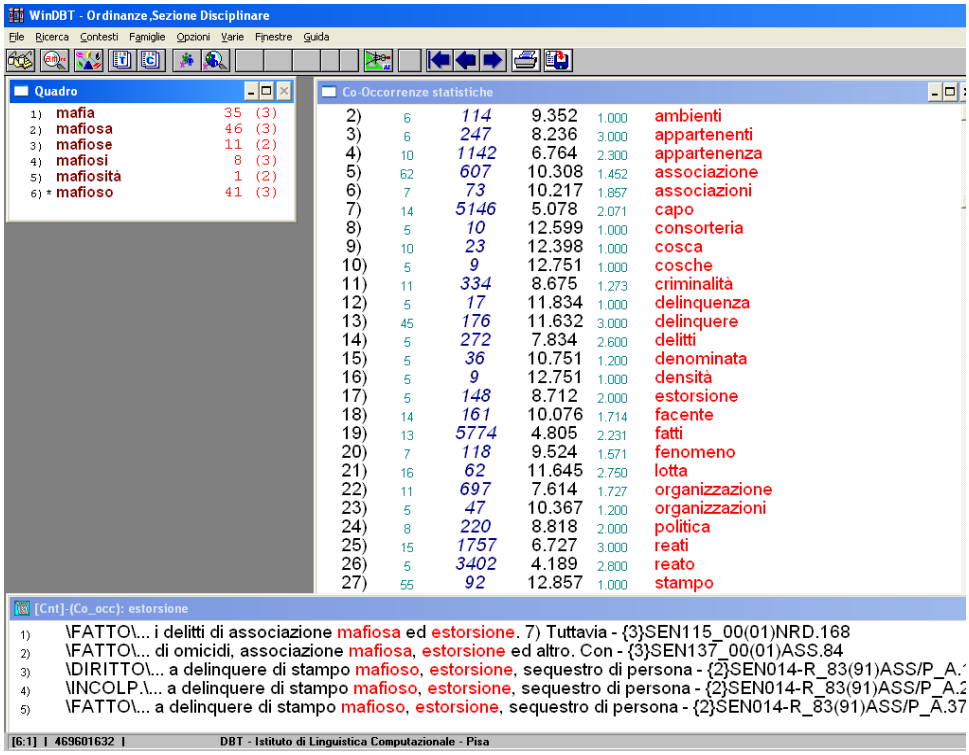
riga, a indicare che quella parola viene ripetuta più volte nella medesima ordinanza:

Mafia - Freq. = 3;
Ord003-x_04(04).455;
Ord003-x_04(04).470;
Ord003-x_04(04).472
Mafiosa - Freq. = 18;
Ord011-S_95(95).73;
Ord004_00(00).365;
Ord004_00(00).581;
Ord002-y_01(01).407;
Ord002-y_01(01).409;
Ord002-y_01(01).413;
Ord002-y_01(01).440;
Ord002-y_01(01).452;
Ord002-y_01(01).458;
Ord002-y_01(01).490;
Ord002-y_01(01).496;
Ord011_02(02).47;
Ord013_02(03).379;
Ord008-z_05(05).542;
Ord008_06(07).53;
Ord004_07(07).62;
Ord004_07(07).145;
Ord004_07(07).183
Mafiosi - Freq. = 3;
Ord004_00(00).220;
Ord002-y_01(01).434;
Ord003-x_04(04).466
Mafioso - Freq. = 4;
Ord003-1_94(98).93;
Ord005_00(00).67;
Ord005_00(00).87;
Ord002-y_01(01).537

4.1.2.Co-occorrenze

La funzione che genera le Co-occorrenze può essere applicata sia nella fase di consultazione della banca-dati a partire da una o più parole, sia sugli archivi in generale.

Figura 5 - Risultato della funzione di Co-occorrenze



Nel caso in oggetto sono state richieste tutte le forme che soddisfino l'etichetta *mafi**. Successivamente viene richiamata la funzione "Co-occorrenze Statistiche" che si basa sulla formula di *mutual information* (Hindle, 1994; definizione a p.123), basata su calcoli statistici della prossimità e della frequenza nei dintorni delle parole cercate, il cui risultato compare nella Figura 5.

Nel riquadro a sinistra compaiono le parole richieste, in quello a destra il risultato, che consiste in un elenco di parole che si trovano in prossimità di quelle di partenza per almeno 5 volte. Nel riquadro in basso è stata cliccata la parola *estorsione* (n.17 della lista) e vengono mostrati i contesti in cui compare in prossimità di alcune delle parole date e a fine riga appare il riferimento logico della sentenza.

4.1.3.Frequenze alfabetiche/decrescenti

Con questa funzione si ottiene l'elenco di tutte le parole della banca-dati o in ordine alfabetico o in ordine decrescente. Si applica una sola volta e può essere archiviata per usi successivi. L'ordine decrescente può essere utile per individuare velocemente i temi chiave di un testo, soprattutto se breve.

Nel caso di un Corpus esteso può essere dispersivo e in questo caso prevale di gran lunga il linguaggio tecnico, mentre potrebbe essere applicato più proficuamente alle singole sentenze. Come esempio si dà un frammento delle parole di senso compiuto più usate (escludendo quindi tutte le forme funzionali) e si elencano in ordine alfabetico per facilitarne la lettura:

4498	appello	12260	art.	5964	atti
4459	attività	4870	avv.	4248	azione
5146	capo	5261	carico	3616	caso
7778	cassazione	4292	componente	2966	comportamento
3371	considerazione	9935	consiglio	11879	corte
11491	data	3792	decisione	3153	depositato
3415	dibattimento	3090	dichiarazioni	21950	disciplinare
43658	dott.	5967	dott.ssa	3896	eletto
5774	fatti	7651	fatto	6488	funzioni
14133	generale	3351	giorni	7354	giudice
5133	giudiziario	3996	giudizio	3717	giustizia
9286	incolpato	6286	incolpazione	3981	indagini
4460	lavoro	3496	legge	3021	luogo
5042	maggio	3785	magistrati	26521	magistrato
6940	magistratura	10678	merito	2948	ministro
3771	nota	3272	oggetto	7601	ordine
3928	parlamento	7646	parte	6490	penale
3355	periodo	13271	presidente	3503	prestigio
3281	procedimenti	12745	procedimento	3499	processo
5425	procura	14794	procuratore	3623	provvedimenti
5036	provvedimento	3860	r.d.l.	3402	reato
4692	relazione	6372	repubblica	6783	richiesta
3274	riferimento	3512	rinvio	3907	ritardi
3767	secondo	4434	sede	3363	segretario
3971	segreteria	3300	seguito	7934	sentenza
5357	sentenze	13874	sezione	3907	sostituto
6979	superiore	11516	tribunale	7103	udienza
8906	ufficio	4310	violazione		

Ci sia permessa una annotazione: in una verifica sulle parole più frequenti precedentemente effettuata nella banca-dati, la frequenza di dott.ssa rappresentava l'11% di quella di dott., mentre invece con l'ultimo aggiornamento (2004-2007) è aumentata al 13,6%.

4.1.4.Indice di leggibilità

Per quanto riguarda l'indice di leggibilità si tratta di una funzione che calcola il valore statistico della *leggibilità* di un testo secondo la formula

di Flesh, riservata alla lingua inglese ma rivista e riadattata all'italiano col nome Gulpease (Lucisano e Piemontese, 1988). I calcoli statistici vengono effettuati sulla base della quantità di frasi (da punto a punto), della quantità di parole e della loro lunghezza. Si tratta quindi di un sistema puramente numerico che ha il fondamento nella maggiore difficoltà di comprensione di frasi o parole molto lunghe.
Nella tabella 8 si riporta una parte del risultato calcolato su alcune sentenze, dove l'indice risulta mediamente basso rispetto allo standard (questo può essere rappresentato da articoli di stampa, dove si attesta tra 46.000 e 64.000).

Tabella 8 - Risultato dell'applicazione dell'indice di leggibilità

Frasi	Parole	Lettere	-LP-	-FR-	*GULPEASE*	Unità di testo
62	2770	15832	571.552	2.238 *	38.560 *	SEN004_90(90)
195	8958	50229	560.717	2.177 *	39.459 *	SEN005_90(91)
83	4671	25345	542.603	1.777 *	40.070 *	SEN006_90(91)
88	4327	23945	553.386	2.034 *	39.763 *	SEN010_90(92)
62	2460	13518	549.512	2.520 *	41.610 *	SEN015_90(90)
77	3879	21322	549.678	1.985 *	39.987 *	SEN016_90(90)
59	3697	21022	568.623	1.596 *	36.925 *	SEN017_90(91)
30	1509	8705	576.872	1.988 *	37.277 *	SEN019_90(90)
32	1439	8162	567.199	2.224 *	38.951 *	SEN025_90(90)
82	3663	20862	569.533	2.239 *	38.762 *	SEN026_90(92)

4.2.Ricerca per Strutture Concettuali

Un'altra applicazione del sistema DBT consente di consultare il testo a partire da strutture concettuali precedentemente impostate e fornite sotto forma di gerarchie di termini significativi.
La struttura deve essere definita anteriormente tramite un generico programma di gestione testi, compreso il blocco-note, purché al momento del salvataggio venga usata l'estensione RSI che permette al DBT di riconoscerla durante la funzione "Ricerca Strutture Concettuali". La sintassi per la stesura di tale struttura prevede alcune regole per la definizione dei livelli e l'inserimento delle condizioni. Trattandosi di questioni abbastanza complesse, dipendendo altresì dalla stretta collaborazione interdisciplinare fra esperti dei settori coinvolti, ci si riserva di approfondire l'argomento, ritenendo importante in questa sede segnalarne le potenzialità.
Lo schema della struttura concettuale qui presentata è stato costruito traendo spunto dalle voci di classificazione utilizzate nel Quaderno N.143

(Sezione Disciplinare del CSM, 2004). La struttura è stata implementata solo per alcune voci, ed è stata applicata ad un sottoinsieme delle sentenze; i numeri tra parentesi graffe indicano la quantità di contesti recuperati da 1471 provvedimenti.

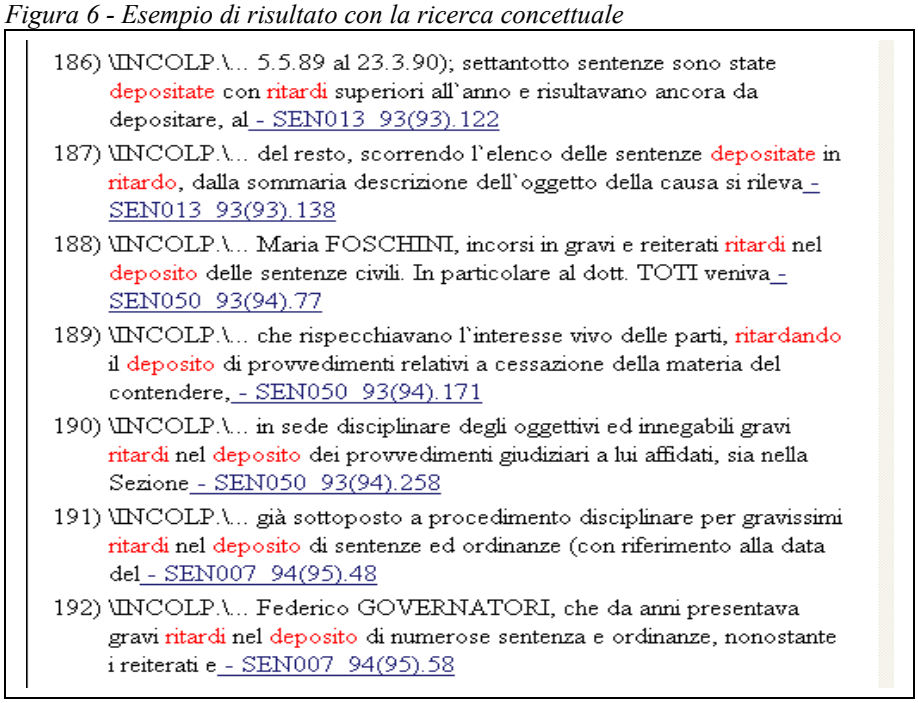
```
(1)-- VOCI DI CLASSIFICAZIONE
(2)--- AZIONE DISCIPLINARE
(3)----- CONDIZIONE
(4)------ [*]APPARTENENZA DELL'INCOLPATO ALL'ORDINE GIUDIZIARIO {344}
          CESSAZIONE & APPARTENENZA {344}
(2)--- CONDOTTA
(3)----- DOVERI DEL MAGISTRATO - CORRETTEZZA
(4)------ [*]VIGILANZA {205}
          VIGILANZA {205}
(3)----- DOVERI DEL MAGISTRATO - IMPARZIALITA'
(4)------ ASTENSIONE
(3)----- DOVERI DEL MAGISTRATO - OPEROSITA'
(4)------ [*]RITARDO O OMISSIONE NELL'ADOZIONE DI PROVVEDIMENTI {305}
           RITARD* & DEPOSIT* {305}
(2)--- [*]SANZIONI {1296}
           FUNZIONI & STIPENDIO {181}
           PERDITA & ANZIANITA {123}
           SANZIONE & AMMONIMENTO {566}
           SANZIONE & CENSURA {368}
           RIMOZIONE {58}
(2)--- SOSPENSIONE CAUTELARE
(3)----- [*]APPLICAZIONE {4}
           ACCOLTA & SOSPENSIONE {4}
(3)----- [*]REVOCA {54}
           REVOCA & SOSPENSIONE {54}
```

Nello schema ogni voce o sottovoce di classificazione è stata "definita" mediante parole o coppie di parole co-occorrenti. Ad esempio sotto la voce SANZIONI, uno dei concetti è stato definito tramite SANZIONE & AMMONIMENTO, che compare 566 volte.

In questo modo tutte le forme rintracciate saranno evidenziate dal sistema e sarà possibile risalire direttamente al contesto. I risultati possono anche essere salvati o stampati. E' possibile recuperare più parole co-occorrenti utilizzando il mascheramento per individuare le forme al plurale, al singolare e quelle di un verbo. Nell'esempio proposto dalla figura 6 è stata scelta la voce RITARDO, che viene espressa con le stringhe RITARD* & DEPOSIT* tramite le quali si reperiscono i contesti di *ritardo*, *ritardare*, *ritardato*, etc. insieme a *deposito*, *depositato*, *depositati*, *depositare*, etc senza vincoli sulla sequenza. Il frammento è stato ritagliato dalla finestra che si apre cliccando sulle singole voci dello schema precedente. Questa opzione può essere utilizzata per rilevare forme alternative che esprimono uno stesso concetto, con il rischio tuttavia che il numero dei contesti

recuperati può aumentare notevolmente, anche se dall'esempio si può notare come tutti i contesti corrispondano al concetto richiesto.

Figura 6 - Esempio di risultato con la ricerca concettuale



4.3.Altri tipi di elaborazioni

Nei paragrafi precedenti sono state descritte ed esemplificate le funzioni del DBT, così come sono disponibili nel pacchetto che viene distribuito nella versione standard (www.ilc.cnr.it/pisystem/prodotti/index.html) in download gratuito. Poiché l'ILC è un istituto di ricerca e sviluppo, vi sono molte altre funzioni che ci permettono di utilizzare la galassia DBT e il PiSystem come componenti in fase di studio e di implementazione.

Si citano ad esempio il Lemmatizzatore o Tagger, che è integrato parzialmente nel sistema di interrogazione, in quanto permette di riconoscere, a livello morfologico, le forme dei lemmi.

E' inoltre possibile applicare all'archivio il riconoscimento delle funzioni svolte dalle parole nella frase (part-of-speach-tagging). In questa ottica sono state utilizzate le procedure, tuttora in fase di sviluppo e di studio, del Linguistic Miner (LM), che è un sistema integrato per la creazione di

grandi archivi testuali di riferimento per la lingua italiana e per l'estrazione rapida di casistiche che puntano alla sintesi della conoscenza linguistica da grandi quantità di dati (Picchi et alii, 2004).

4.3.1.Ricerca di schemi linguistici

Nella sperimentazione del LM sono stati utilizzati i testi indicizzati tramite la funzione di tagging automatico, che assegna a ciascuna parola il suo lemma, la categoria grammaticale e la sua funzione nella frase.
In seguito sono state applicati ai testi lemmatizzati dei criteri di ricerca espressi attraverso schemi linguistici: tramite una sintassi data vengono poste condizioni sulle categorie grammaticali e vengono così estratte sequenze di parole che rispettano tali regole.
Nell'esempio che segue sono state estratte le coppie di sostantivi, con eventuali aggettivi, connesse da preposizione (semplice o articolata); il risultato di questa ricerca è stato poi ordinato in base alla frequenza in modo da individuare facilmente le "formule" più frequenti nel linguaggio dei provvedimenti.
Dai risultati ottenuti è stata ottenuta la tabella 6, in cui sono elencate, in doppia colonna e ciascuna con la frequenza corrispondente, le "formule" precedentemente individuate e poste in ordine alfabetico per una migliore lettura. Vi sono casi di varianti con frequenza più bassa, che per questo motivo non compaiono in questa lista, come ad esempio *Procuratore generale*, che viene usato anche sotto forma di sigla [P.g.].

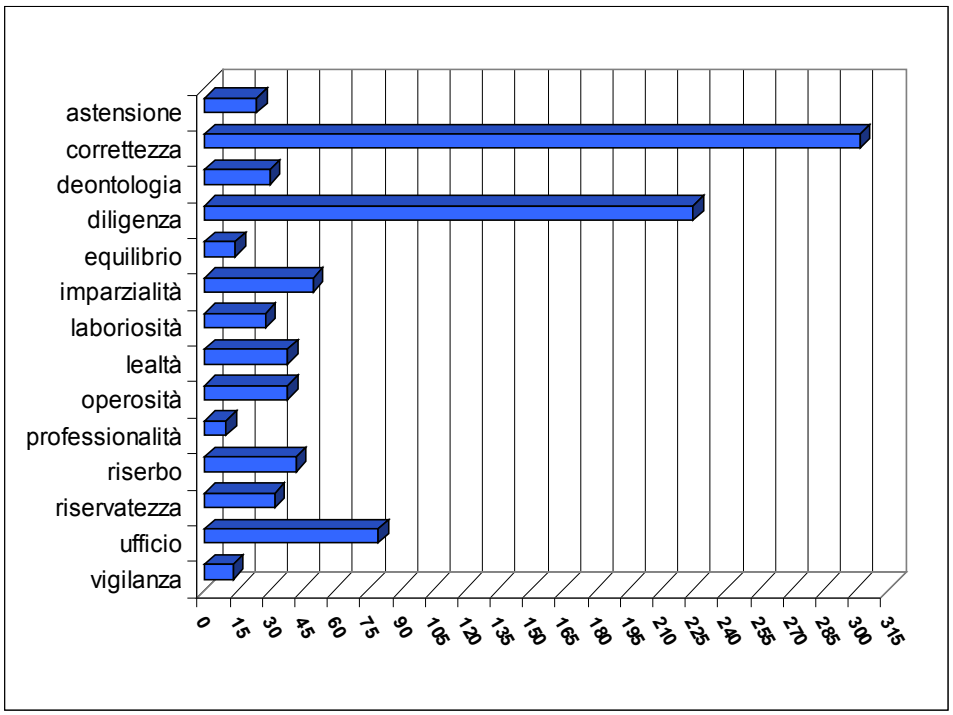
Tabella 6 - Risultato della ricerca per schema linguistico

Freq.	*Frase*	*Freq.*	*Frase*
341	ambito del procedimento	195	motivazione della sentenza
186	amministrazione della giustizia	1362	motivi della decisione
160	archiviazione del procedimento	1883	nome del popolo
1081	assistenza del segretario	188	ordine degli avvocati
327	atti del procedimento	255	presidente del collegio
428	capo dell' ufficio	1887	presidente del consiglio
1240	carico del dott.	1857	presidente del tribunale
402	cessazione dell' appartenenza	1151	presidente della corte
188	comitato dei creditori	633	presidente della sezione
267	comportamento del dott.	2009	prestigio dell' ordine
257	compromissione del prestigio	362	prestigio delle funzioni
200	confronti dei dottori	1448	procura della repubblica
3499	confronti del dott.	3350	procuratore della repubblica
168	confronti del dottor	1006	procuratore generale della repubblica
162	confronti del magistrato	257	richiesta del procuratore

503	confronti della dott.ssa	291	sanzione dell' ammonimento
518	consigliere della corte	196	sanzione della censura
248	consiglio dell' ordine	349	sanzione disciplinare dell' ammonimento
8303	consiglio superiore della magistratura	235	sanzione disciplinare della censura
278	corso dell' istruttoria	167	scadenza dei termini
244	credibilità dell' ordine	219	scadenza del termine
311	deposito dei provvedimenti	1147	segreteria del consiglio
621	deposito delle sentenze	1161	sensi dell' art.
565	direttore della segreteria	278	sentenza della corte
163	discussione orale del procedimento	231	sezione del tribunale
233	epoca dei fatti	165	sezione della corte
256	esercizio dell' azione	346	sezione disciplinare del c.s.m.
157	esercizio della funzione	4574	sezione disciplinare del consiglio
879	esercizio delle funzioni	230	sezione disciplinare del csm
322	esito dell' istruttoria	161	sospensione del procedimento
208	fini della nomina	405	sostituzione del vice
263	fissazione dell' udienza	706	svolgimento del procedimento
214	g.i.p. del tribunale	765	svolgimento del processo
918	giudice del tribunale	314	titolari dell' azione
1135	intervento del sostituto procuratore	200	trasmissione degli atti
190	irrogazione della sanzione	367	tribunale del riesame
176	lesione del prestigio	388	violazione dei doveri
188	ministero della giustizia	323	violazione del dovere
1032	ministro della giustizia	2025	violazione dell' art.

Prendendo spunto dall'alta frequenza della frase "violazione(i) del(i) dovere(i) ...", si è analizzato nel dettaglio il contesto seguente e nella figura 5 viene mostrata una sintesi grafica del suo utilizzo nella banca-dati, mostrando gli usi più frequenti di 10 volte e riconducendo i sinonimi al lemma di base.

Figura 5 Sintesi grafica dell'uso della frase "violazione del dovere ..."



4.4.Sperimentazione dell'anonimazione di testi

E' stata sperimentata l'anonimazione dei testi dell'archivio delle Ordinanze della Sezione Disciplinare del CSM, in modo da rendere disponibile il database privato dei dati personali. Il riconoscimento dei Nomi Propri nei testi avviene tramite l'utilizzo di un componente del Linguistic Miner, ossia la funzione denominata CercaFenomeni.

Questa operazione comporta varie difficoltà, poiché tali entità (le cosiddette Named Entity) possono essere costituite dal solo nome, cognome o da entrambi in vario ordine o abbinati a iniziali. Una volta individuate tali entità, con l'ausilio di un elenco di nomi già catalogati, queste vengono memorizzate in un file su cui è possibile intervenire manualmente per apportare modifiche come inserimento o esclusione di elementi. Successivamente viene generato un nuovo archivio, in cui ciascun elemento della lista precedente viene marcato come nome proprio. Quindi nei provvedimenti disciplinari le entità marcate sono sostituite da una stringa alfanumerica composta di 2 caratteri tra *, in modo tale che

nell'ambito di una stessa Ordinanza ogni occorrenza del nome e/o cognome, riferito alla medesima persona, sia sostituita con uguale sigla.
La verifica dei risultati della prova effettuata sull'archivio delle Ordinanze ha evidenziato ulteriori problematiche: la difficoltà di distinguere i nomi di persona da quelli dei luoghi, che non devono essere anonimati. Sebbene tale inconveniente possa essere risolto con un dizionario dei nomi geografici, in fase di costruzione all'interno del progetto LM, tuttavia in alcuni casi permane l'ambiguità; ad esempio Rossano si riferisce talvolta ad una località, talvolta costituisce nome di persona, che se non associato al relativo cognome o nome, può essere disambiguato solamente valutando il contesto.

5. Sviluppi futuri e conclusioni

L'archivio dei provvedimenti descritto costituisce una banca dati con cui monitorare quasi venti anni di lavoro della sezione disciplinare. Nello stesso tempo costituisce un nucleo che può essere integrato sia con gli anni precedenti (anche utilizzando le pubblicazioni già effettuate) sia con l'aggiornamento dei provvedimenti che man mano vengono emessi. Il lavoro di analisi già fatto semplifica la codifica dei nuovi documenti.
Tramite le codifiche inserite sono possibili anche altri tipi di analisi mirate, effettuando estrazioni e creando dei sotto-archivi di sentenze/ordinanze omogenei, ad esempio, per sanzioni comminate o per sedi territoriali, ecc.
Inoltre occorre segnalare che, trattandosi di dati testuali già codificati nei fenomeni più rilevanti e già strutturati organicamente, è possibile passare ad altre codifiche in modo da poter trasmigrare i dati ad altre applicazioni che sono già in uso presso le istituzioni competenti. Quest'ultima considerazione è in linea con la tendenza della Pubblica Amministrazione che consiglia il "riuso" di strumenti e dati già elaborati.
Siamo consapevoli che certe funzioni, messe a punto principalmente per scopi di analisi linguistica, possano apparire troppo specialistiche in un settore dove altri sono gli obiettivi; comunque riteniamo che tali strumenti, opportunamente rivisitati con la stretta collaborazione degli specialisti, possano veramente offrire delle potenzialità ancora inesplorate.

6. Bibliografia

Ceccotti, M.L. e Sassi, M. (2003), *L'archivio elettronico delle opere di Carlo Emilio Gadda. Da redattori a fruitori di un data base testuale*, in Linguistica Computazionale, Pisa-Roma, Istituti Editoriali e Poligrafici Internazionali (IEPI), pp. 221-250.

Cinini, A. e Sassi, M. (2004), *Attività nell'ambito del WP1: Prototipo di analisi testuale a supporto del WP2*, IRSIG-ILC-CNR, Pisa.

Hindle, D. (1994) *A Parser for Text Corpora* in Atkins, B.T.S., Zampolli, A. (eds.) Computational Approaches to te Lexicon, Clarendon Press, Oxford, pp.103-151.

Lucisano, P. e Piemontese, E. (1988), *GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana,* in Scuola e città, n.3, Firenze, La Nuova Italia Editrice, pp. 110-124.

Picchi, E. (1997), *DBT 3 - Data Base Testuale: Guida all'uso*, ver. 3.1. Lexis Ricerche s.r.l. su licenza del C.N.R., Roma.

Picchi, E. (1999), *Informatica e scienze umane: Procedure di analisi testuale* in Zanetti, M.A. (a cura di), Parola e Immagine, Firenze, La Nuova Italia Editrice.

Picchi, E., Ceccotti, M.L., Cucurullo, S., Sassi, M. e Sassolini, E. (2004), *Linguistic Miner: an Italian Linguistic Knowledge System* in LREC 2004 IV International Conference, Proceedings, Lisbona, 26-28 maggio 2004, pp.1811-14.

Sassi, M. (2003), *La consultazione dei corpora costituzionali con il DBT*, in Diritto alla vita e Diritto all'ambiente nel lessico costituzionale e nella dottrina giuridica. Strumenti e metodi per l'analisi linguistica-concettuale, con CD-Rom allegato, Firenze, S.T.A.R., pp. 149-154.

Sassi, M. e Cinini, A. (2006), *Content-analysis dei provvedimenti della sezione disciplinare del C.S.M.*, ILC-IRSIG-CNR, Pisa, S.T.A.R.

Sezione Disciplinare del Consiglio Superiore della Magistratura, (2004) *Massimario delle decisioni, Anni 2002-2003*, Quaderno n.143 (http://www.csm.it/pages/quaderni.html).