DOI: 10.17469/O2106SLI000012

FABIO TAMBURINI

I corpora del FICLIT, Università di Bologna: CORIS/CODIS, BoLC e DiaCORIS

Presenteremo i tre grandi corpora sviluppati dal nostro gruppo di ricerca e liberamente consultabili sul Web.
Il progetto più rilevante ha riguardato la creazione di CORIS, un corpus generale rappresentativo dell'italiano contemporaneo; è disponibile online dal 2001 e costantemente aggiornato, ogni tre anni, per cogliere le variazioni nel lessico e nei rapporti di frequenza. CODIS, la sua controparte dinamica, consente allo studioso di combinare liberamente insiemi predefiniti di testi per costruire il corpus ideale per studi specifici o per studi interlinguistici.
BoLC è un corpus bilingue (italiano/inglese) di linguaggio giuridico contenente una sezione parallela e una sezione comparabile. Consente lo studio e il confronto della terminologia adottata nei due sistemi giuridici ed è disponibile online dal 2011.
DiaCORIS, online dal 2006, riproduce la struttura in macro-varietà testuali di CORIS in una prospettiva diacronica. Contiene testi dall'Unità d'Italia agli anni 2000 adeguatamente annotati con i principali metadati.

Parole chiave: corpora, rappresentatività, struttura dinamica.

1. *Introduzione*

Questo breve contributo intende presentare i grandi corpora che sono stati sviluppati al Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna.

Negli ultimi 25 anni abbiamo sviluppato tre grandi corpora: il primo, e più importante progetto, riguarda CORIS/CODIS, il corpus di riferimento per l'italiano scritto contemporaneo; il secondo corpus che abbiamo sviluppato è un corpus terminologico, bilingue, contenente linguaggio giuridico e composto da una sezione in lingua inglese e una sezione in italiano; il terzo Corpus, DiaCORIS, è un corpus diacronico di riferimento per l'italiano dall'Unità d'Italia al 2001.

2. *CORIS/CODIS*

CORIS è il progetto decisamente più complesso che abbiamo affrontato (Rossini Favretti *et al*. 2002). È configurato come un COrpus di Riferimento per l'Italiano Scritto contemporaneo, è stato sviluppato a partire dalla fine degli anni '90 e attualmente, con l'ultimo aggiornamento inserito nell'estate 2021, contiene circa 165 milioni di parole, o meglio di *token*. CORIS viene aggiornato ogni tre anni mediante un monitor corpus in modo da mantenere il lessico e i rapporti di frequenza tra le varie aree semantiche aggiornati e di cogliere tutte le evoluzioni storiche avvenute in questi ultimi anni.

Come possiamo vedere nella Tabella 1, CORIS è costituito da una prima porzione di testi dagli anni '80 al 2000 contenente circa 100 milioni di parole ed è stato regolarmente aggiornato ogni tre anni inserendo un monitor corpus di 10 milioni di parole al fine di catturare i grandi eventi degli ultimi vent'anni.

Tabella 1 - *Struttura delle sezioni temporali di CORIS/CODIS e alcuni eventi storici rilevanti considerati nei vari monitor corpora*

Sezione temporale	*Dim.*	*Descrizione/Eventi*
CORIS 1980-2000:	100Mw	◃ Sezione iniziale di CORIS
Monitor 2001-04:	10Mw	◃ 11 settembre 2001, Euro, Guerre in Medio Oriente
Monitor 2005-07:	"	◃ Italia Camp. Mondo di Calcio
Monitor 2008-10:	"	◃ Crisi economica globale
Monitor 2011-13:	"	◃ Incidente di Fukushima
Monitor 2014-16:	"	◃ Fond. Islamico, Terrorismo
Monitor 2017-20:	14Mw	◃ Primo anno di pandemia

CORIS è disponibile online dal 2001 per tutti gli studiosi e gli studenti: si configura quindi come un corpus di riferimento generale sincronico, contiene unicamente lingua scritta, testi autentici e integrali ed è annotato automaticamente rispetto alle categorie grammaticali (PoS-tag) e ai lemmi.

Per definire la struttura gerarchica del corpus (Biber 1993), ci siamo attenuti rigorosamente a criteri esterni (Atkins *et al*. 1992); per il bilanciamento dei vari subcorpora abbiamo considerato in prima istanza parametri quantitativi relativi alla circolazione dei quotidiani e alla distribuzione dei volumi, modulati, in una seconda fase, da parametri qualitativi, per riproporzionare al meglio i rapporti tra diverse

tipologie testuali, come ad esempio il livello di attenzione cognitiva e il tipo di uso dei testi che compongono il corpus.

CORIS e tutti gli aggiornamenti successivi rispettano quindi le proporzioni indicate nella Tabella 2 tra le sei principali macro-varietà testuali considerate.

Tabella 2 - *Tipologia e proporzioni tra le principali macro-varietà testuali (subcorpora) di CORIS/CODIS*

Subcorpus	*Proporzione*
Stampa	38%
Narrativa	25%
Prosa Accademica	12%
Prosa Giuridico-Amministrativa	10%
Miscellanea	10%
Ephemera	5%

Accanto a CORIS, ovvero composto dagli stessi documenti ma strutturato in modo completamente diverso, abbiamo introdotto CODIS, il COrpus Dinamico dell'Italiano Scritto (Tamburini 2002). Due sono state le ragioni che ci hanno spinto a sviluppare una versione dinamica di CORIS: innanzitutto, il nostro studio di rappresentatività, per quanto molto accurato, potrebbe non essere condiviso da altri studiosi e, in seconda istanza, per consentire studi interlinguistici basati su corpora. Come si può osservare dalla Tabella 3 (un estratto da Tamburini 2002) la struttura dei vari corpora nel panorama internazionale mostra una rilevante variazione nelle tipologie testuali e nelle proporzioni.

Tabella 3 - *Struttura e bilanciamento di alcuni corpora nel panorama internazionale (valori riferiti al 2002)*

Corpus	*Composizione*	
BNC	Volumi	52.5 Mw – 58.6%
90Mw – Inglese	Stampa	27.8 Mw – 31%
(solo lingua scritta)	Miscellanea	7.4 Mw – 8.3%
LSWE	Fiction	5 Mw – 17.8%
28Mw – English	News	10.6 Mw – 37.7%
(solo lingua scritta)	Prosa Accademica	5.3 Mw – 19%
	Prosa generica	6.9 Mw – 24.6%

Corpus	*Composizione*	
The Oslo Corpus	Fiction	3.8 Mw – 17%
22.3 Mw – Norvegese	Quotidiani/Riviste	10.6 Mw – 47.5%
	Prosa	7.8 Mw – 35%
Corpus de Referência do	Quotidiani	55 Mw – 60.8%
Português Contemporâneo	Volumi	20.5 Mw – 22.6%
(CRPC) – 92 Mw	Periodici	7 Mw – 7.7%
Portoghese	Decisioni della	1.8 Mw – 2%
(solo lingua scritta)	Suprema Corte di Giustizia	
	Miscellanea	3.9 Mw – 4.3%
	Opuscoli	0.3 Mw – 0.3%
	Lettere	0.1 Mw – 0.1%

La struttura dinamica di CODIS consente allo studioso di selezionare quali porzioni dei materiali di CORIS considerare nelle ricerche. I testi sono esattamente gli stessi di CORIS e sono stati semplicemente raggruppati in "sezioni" di svariati milioni di parole (si veda la Tabella 4): ogni subcorpus è suddiviso in quattro sezioni di materiali testuali con proporzioni differenti e lo studioso/studente può comporre il corpus da utilizzare per le indagini rispetto alle proprie sensibilità o le proprie necessità. Combinando queste sezioni liberamente, CODIS consente studi interlinguistici basati su corpora o consente di lavorare con un corpus strutturato diversamente o pensato con differenti criteri di rappresentatività rispetto alla progettazione iniziale.

Tabella 4 - *Dimensione delle sezioni di materiali selezionabili in CODIS rispetto ai vari subcorpora*

Subcorpus	*Dimensioni (in %)*			
Stampa	20	10	5	3
Narrativa	13	7	3	2
Prosa Accademica	5	4	2	1
Prosa Giuridico-Amministrativa	4	3	2	1
Miscellanea	4	3	2	1
Ephemera	2	1	1	1

3. *BoLC*

Il secondo grande progetto riguarda il *Bononia Legal Corpus* (Rossini Favretti *et al*. 2007). BoLC è un corpus di linguaggio giuridico bilin-

gue, contiene testi in lingua inglese britannica e in lingua italiana ed è stato strutturato in due sezioni distinte: una prima sezione contiene documenti paralleli, cioè in rapporto di traduzione, ed è basata prevalentemente su documenti dell'Unione Europea (direttive e sentenze); la seconda sezione contiene un corpus comparabile contenente testi estratti dalla giurisprudenza e dalla legislazione dei due paesi. La Tabella 5 mostra le tipologie di testi introdotte nelle due sezioni.

In questa duplice prospettiva, si è prevista la possibilità di giungere a una comparazione di testi giuridici, le cui forme sono espressione, da un lato, di un ordinamento giuridico comune, quale quello comunitario, e, dall'altro, di ordinamenti e culture giuridiche diverse, quali quelle sviluppate nei singoli stati. Si è inteso, in tal modo, tenere conto sia la progressiva formazione di un diritto uniforme a livello europeo sia la pluralità di sistemi normativi nazionali nell'ambito dell'Unione Europea.

Tabella 5 - *La struttura del corpus BoLC*

	Sezione 1	*Sezione 2*
Inglese	Directives, Judgments	Acts of Parliament, Chancery Division, Court of Appeal, Family Division, House of Lords, Privy Council, Queen's Bench Division, Statutory Instruments
Italiano	Direttive, Sentenze	Costituzione, Codice Civile, Codice Penale, Codice di Procedura Civile, Codice di Procedura Penale, Decreti Legislativi, Leggi Costituzionali, Leggi Ordinarie, Sentenze Penali Corte di Cassazione, Sentenze Civili Corte di Cassazione, Sentenze e Ordinanze della Consulta

4. *DiaCORIS*

L'ultimo grande progetto che presenteremo riguarda DiaCORIS, il corpus diacronico di italiano scritto sviluppato in collaborazione con l'Accademia della Crusca e L'Università di Modena e Reggio Emilia (Onelli *et al*. 2006). L'obiettivo di DiaCORIS è quello di integrare CORIS in una prospettiva diacronica: contiene testi a partire dall'Unità d'Italia nel 1861 fino al 2001, esattamente in corrispondenza dell'inizio temporale dei monitor corpora di CORIS, costituendo quindi un continuum temporale che copre l'evoluzione storica della lingua italiana dall'Unità fino a oggi. L'intervallo temporale è stato suddiviso in cinque sezioni composte da 5 milioni di parole ciascuna

per un totale di 25 milioni di parole. Ogni sezione ricalca la struttura delle macro-varietà testuali di CORIS (eccetto l'Ephemera, non presente in diacronia); le proporzioni tra le varie sezioni cambiano nel tempo in funzione dell'importanza della macro-varietà testuale nel periodo storico considerato (si veda la Tabella 6).

Tutti i testi di DiaCORIS contengono i metadati fondamentali che li caratterizzano (si veda la Tabella 7) in modo da consentire studi storici accurati e completi.

Tabella 6 - *Suddivisione temporale e proporzione tra le varie sezioni del corpus DiaCORIS*

Sezione	*Subcorpus*	*Prop.*
1861-1900 Dopo l'unificazione	STAMPA	15%
	NARRATIVA	30%
	SAGGISTICA	30%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	15%
1901-1922 Il Periodo Liberale	STAMPA	25%
	NARRATIVA	25%
	SAGGISTICA	25%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	15%
1923-1945 Periodo Fascista	STAMPA	30%
	NARRATIVA	25%
	SAGGISTICA	25%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	10%
1946-1967 Dopo la 2a Guerra Mondiale	STAMPA	35%
	NARRATIVA	20%
	SAGGISTICA	25%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	10%
1968-2001 Dopo la Rivoluzione del '68	STAMPA	40%
	NARRATIVA	20%
	1968 SAGGISTICA	20%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	10%

Tabella 7 - *Struttura dei metadati inseriti nei testi che compongono DiaCORIS in formato XML*

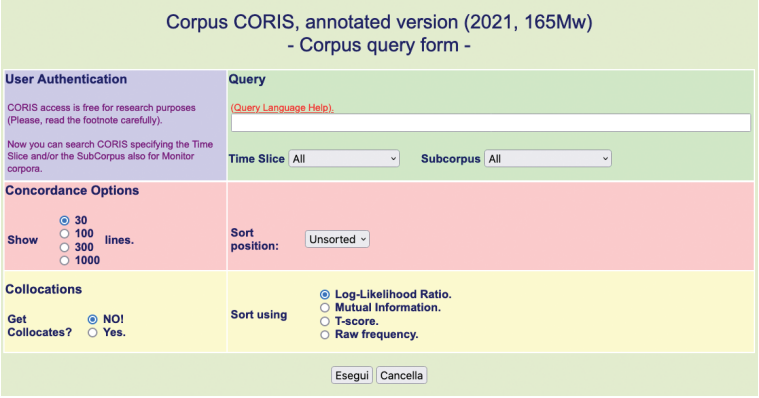
```
<text id="filename.xml" lingua="Italiano"
      titolo="titolo del testo"
      autore="nome autore" ed_test="editore/testata"
      anno="anno di pubblicazione"
      sez="sezione" subc="subcorpus">
word1
word2
...
</text>
```

5. *Interfaccia di consultazione dei corpora*

La Figura 1 mostra l'interfaccia di consultazione di tutti i corpora. Consente di effettuare tutte le più comuni operazioni per l'estrazione di informazioni da corpora elettronici (Sinclair 1991; 2004): l'estrazione delle concordanze di un termine, eventualmente incrociando le informazioni lessicali con le annotazioni, l'estrazione delle collocazioni di un nodo, avvalendosi di vari indici statistici di associazione, e l'estrazione di informazioni sulla frequenza dei termini.

L'accesso ai corpora è completamente libero all'indirizzo https://corpora.ficlit.unibo.it.

Figura 1 - *L'interfaccia di consultazione di CORIS, comune a tutti i corpora del FICLIT*



Riferimenti bibliografici

Atkins, Sue & Clear, Jeremy & Ostler, Nicholas. 1992. Corpus Design Criteria, *Literary and Linguistic Computing*, 7(1), 1-16.

Biber, Douglas. 1993. Representativeness in corpus design. *Journal of Literary and Linguistic Computing*, 8(4), 243-257.

Onelli, Corinna & Proietti, Domenico & Seidenari, Corrado & Tamburini, Fabio. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proc. 5th International Conference on Language Resources and Evaluation – LREC 2006*, 1212-1215. Genova.

Rossini Favretti, Rema & Tamburini, Fabio & De Santis, Cristiana. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Wilson, Andrew & Rayson, Paul & McEnery, Tony (a cura di), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, 27-38. Lincom-Europa, Munich.

Rossini Favretti Rema, Tamburini Fabio & Martelli Edoardo. 2007. Words from Bononia Legal Corpus. In W. Teubert (a cura di), *Text Corpora and Multilingual Lexicography*, John Benjamins Publishing Company, 11-30.

Sinclair, John. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2004. Carter, Ronald (a cura di), *Trust the Text: Language, Corpus and Discourse*. Routledge.

Tamburini, Fabio. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rossini Favretti, Rema (a cura di), *Linguistica e*

informatica: multimedialità, corpora e percorsi di apprendimento, 57-73. Bulzoni, Roma.

Tamburini, Fabio. 2002. A dynamic model for reference corpora structure definition. In *Proc. Third International Conference on Language Resources and Evaluation – LREC2002*, 1847-1850. Las Palmas, Canary Islands, Spain.

Tamburini, Fabio. 2007. CORISTagger: a high-performance PoS tagger for Italian. *Intelligenza Artificiale*, IV(2). 14-15.