4. Conclusion

To summarise, I would like to point out that the four issues that I have raised are interrelated quite intricately. Unless we are prepared to welcome very large corpora, we will not get access to the information that we need about the languages in order to pick up the challenges of information retrieval. As long as we rely on tags we are forcing the attention (and the resources) on pre-corpus models of language which require only small corpora anyway. Tagged corpora will not meet the requirements of the information society because they are not sensitive enough, and they would have done so by now, since they have had all the attention so far. They have proved particularly unsuccessful with open text, which is an essential part of whatever prescription will be set down for programs that understand human language.

Corpus-driven linguistics demands extremely large corpora because of its need for multiple occurrences of all the items it handles; it rejects manual tagging and invites a complete rethinking of the methods of on-line analysis; it opens up new avenues of research which may help with information retrieval and other applications, and it may get closer to the goal of the machine understanding of language, though extensive testing is recommended before assaults on this goal are attempted.

References

BNC Sampler (1999), release 1.1 (CD-ROM): Oxford University Humanities Computing Unit.

Hunston S. (1993), "Evaluation and ideology in scientific writing", in Ghadessy M. (ed.), *Register Analysis: Theory and Practice*. London, Pinter, pp. 57-74.

Sinclair J. (1999), "New roles for Language Centres: the mayonnaise problem", in Bickerton D. / Gotti M. (eds.), *Language Centres: Integration through Innovation,* CercleS (Conféderation Européenne des Centres de Langues de l'Enseignement Superieur) Secretariat, Department of Modern Languages, University of Plymouth, pp. 31-50.

Sinclair J. (2000), "The Deification of Information", in Thompson G. / Scott M. (eds.), *Patterns of Text: in honour of Michael Hoey*, Amsterdam/Philadelphia John Benjamins, forthcoming.

Somers H. (1997), "A Practical Approach to Using Machine Translation Software – 'Post-editing' the Source Text", in *The Translator*, 3: 2, 193-212.

Tognini Bonelli E. (2000), *Corpus Linguistics at Work*, Amsterdam and Philadelphia, John Benajmin, forthcoming.

Zipf G. (1935), *The Psychobiology of Language*, Houghton Mifflin. Reprinted 1965 Boston, MIT Press.

PROGETTAZIONE E COSTRUZIONE DI UN *CORPUS* DI ITALIANO SCRITTO: CORIS/CODIS

REMA ROSSINI FAVRETTI
Università di Bologna
Centro Interfacoltà di Linguistica Teorica e Applicata "L. Heilmann"

1. Introduzione

Gli studi condotti, negli anni più recenti, nell'ambito della *corpus linguistics* hanno ampiamente mostrato le possibilità di analisi che si aprono a livello metodologico e teorico tramite questo approccio che, profondamente radicato nella tradizione linguistica strutturale, viene a riproporre alcune problematiche tradizionali per considerarle nella nuova luce data dallo sviluppo delle tecnologie sia *hardware* che *software*. La potenza delle attrezzature e dei programmi ci consente di avvicinarci ai dati linguistici presenti nella *langue* con nuove prospettive di indagine.

La massa dei dati elaborabili dal computer e la brevità dei tempi di interrogazione costituiscono il punto focale della *corpus linguistics*, o linguistica dei *corpora*, secondo la formulazione che si è considerata preferibile nella nostra lingua, pur nell'incertezza che sembra gravare in Italia su questo ambito di indagine: un'incertezza a livello linguistico che viene in qualche modo a riflettere una difficoltà riscontrabile a livello epistemologico nella definizione dei tratti e degli aspetti che caratterizzano la linguistica dei *corpora* rispetto ad ambiti disciplinari affini, come, ad esempio, la linguistica computazionale. È lecito supporre ed auspicare che questa disciplina possa in tempi brevi affermarsi nel nostro paese, anche in considerazione del fervore di iniziative e di studi presente in ambito europeo, indicativo di una linea di tendenza che, delineatasi con il primo affermarsi delle tecnologie informatiche, si è poi incessantemente rafforzata, fino ad acquisire, in particolare nell'ultimo decennio, una collocazione propria e definita nella ricerca linguistica.

A questi studi, pur nelle inevitabili differenziazioni, si ricollega il lavoro che ha portato alla costruzione di un *corpus* di italiano scritto, in fase di avanzata realizzazione presso il CILTA, e da me progettato e diretto. Per un'adeguata descrizione occorrerebbe ripercorrere le varie fasi della realizzazione, evidenziando i propositi e le riflessioni che ne hanno determinato la progettazione. Occorrerebbe esporre i criteri informativi e le motivazioni soggiacenti.

Conto di poterlo presto fare in altra sede, mi limiterò qui a descriverne la costruzione considerandone alcune fasi principali quali

1. la progettazione
   a) tipologia del *corpus*
   b) dimensione
   c) rappresentatività
2. l'elaborazione del modello di costruzione
   a) identificazione della popolazione
   b) definizione dei criteri di selezione
3. la definizione della strutturazione del *corpus*
   a) articolazione dei componenti
   b) definizione dei rapporti fra i componenti
   c) campionamento
4. la definizione (del CORpus di Italiano Scritto - CORIS/CODIS)
5. il reperimento e l'inserimento dei materiali
6. la lemmatizzazione e l'annotazione grammaticale.

Nei prossimi paragrafi descriverò la progettualità soggiacente alla definizione del *corpus* considerando i vari punti[1] nelle loro interrelazioni.

2. Descrizione del progetto

Consideriamo in primo luogo che cosa si intende quando si parla di *corpus*. Il concetto di *corpus* ha una lunga tradizione negli studi linguistici che non occorre qui ripercorrere. Appare importante, piuttosto, rilevare come, anche limitandoci all'ambito della recente linguistica dei *corpora,* non si riscontri uniformità nella definizione. Varie sono le definizioni ricorrenti che testimoniano, nella loro diversa articolazione, le linee che si sono privilegiate negli studi condotti nell'ambito della linguistica dei *corpora*.

Nella definizione di Francis il *corpus* è visto come

> a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis (1982, 7)

[1] Il punto 6 costituirà l'oggetto del contributo di F. Tamburini, mio principale collaboratore nell'elaborazione del progetto. Alcuni aspetti del reperimento del materiale saranno esposti da C. De Santis.

Alcuni elementi della definizione possono essere ancora recepiti. Il *corpus* come *definiendum* trova la propria descrizione nella "raccolta di testi", che è posta come *definiens* anche in una definizione data da Sinclair,

> a collection of naturally-occurring language text, chosen to characterize a state or variety of a language (1991, 71)

così come in quella data da Biber *et al.*, secondo cui

> a corpus is a large and principled collection of natural texts (1998, 12)

La "raccolta di testi", o di frammenti di testi, può essere vista come classe di riferimento, ma in nessun caso questa è posta come fine a se stessa o mero accumulo di dati. Nelle diverse definizioni prese in esame il sintagma risulta qualificato e specificato da tratti definitori che, pur differenziandosi, si collocano in un apparato concettuale comune. I testi, o i frammenti, devono essere

a) linguistici
b) autentici e ricorrenti nell'uso
c) in formato elettronico
d) rappresentativi.

Tali tratti vengono ad integrarsi reciprocamente e possono contribuire a configurare una definizione di *corpus* quale

> una raccolta di testi, autentici e ricorrenti nell'uso, in formato elettronico, rappresentativi di uno stato o di una varietà di una lingua.

Se questa definizione può essere accettata in una fase preliminare del lavoro, accogliendo l'apparato concettuale di riferimento che da questa viene offerto, essa appare suscettibile di ulteriori considerazioni e precisazioni in particolare per quanto concerne i termini "testi" e "rappresentativi", per i diversi valori da questi assunti nei vari studi.

Se troviamo accordo e consenso fra gli studiosi sulla procedura di selezione che è operata nella costituzione del *corpus* così come sul carattere di autenticità dei testi raccolti, permangono ambiguità sulla nozione di "testo" che viene recepita ai fini della selezione e, in particolare, sui criteri sottesi alla definizione della rappresentatività dei *corpora.* Si pongono domande a livello sia di progettazione del *corpus* sia di elaborazione del modello di costruzione.

Quali sono i criteri che sottendono alla selezione ed all'assemblaggio dei testi e che consentono di definire l'insieme costituito "una raccolta"? Devono questi essere espliciti o possono essere assunti implicitamente come posto nella definizione di Francis? Devono i testi essere frammentati in sequenze opportunamen-

te dimensionate oppure essere memorizzati nella loro totalità? Quali rapporti devono stabilirsi fra gli insiemi e i sottoinsiemi affinché un *corpus* possa essere considerato rappresentativo? In che misura l'inevitabile soggettività della selezione può venire controbilanciata dalla dimensione dei *corpora* attualmente consentita dallo sviluppo delle tecnologie informatiche? E ancora, in che misura i testi selezionati devono essere rappresentativi dell'insieme o sottoinsieme da cui sono tratti? Torneremo sui vari punti successivamente, ma mi sembra opportuno porre in rilievo fin d'ora questi problemi preliminari che il linguista deve affrontare nella fase di costituzione di un *corpus*, e che, a mio avviso, assumono una crescente rilevanza in relazione alla maggiore potenza delle attrezzature informatiche.

2.1. TIPOLOGIA

Ai fini della progettazione e della costruzione del *corpus* alcune scelte sono state preliminari ponendo la base per le operazioni successive. In primo luogo si è trattato di definire la finalità del progetto e la tipologia del *corpus* che si intendeva costruire.

All'interno dell'ampia tipologia di *corpora* ormai istituita si identificano

— *corpora* generali (*general corpora*) e specialistici (*special/specialized corpora*)
— *corpora* di riferimento (*reference corpora*)
— *corpora* di archivi (*archives corpora*)
— *corpora* di testi interi (*text corpora* o *full-text corpora*) o di campioni di testi (*samples corpora*)[2].

Fin dalle prime fasi della progettazione, si è identificata la finalità del lavoro nella costruzione di un *corpus* generale, per la cui descrizione si poteva ancora fare riferimento alla definizione data del Brown Corpus, uno dei primi *corpora* elettronici. Come il Brown Corpus era stato indicato quale "a standard sample of present-day English for use with digital computers", la finalità della ricerca poteva identificarsi nella costituzione di un insieme di testi informatici rappresentativi, in senso lato, dell'italiano attuale[3]. Nell'identificazione di tale finalità trovava risposta uno dei primi problemi che si ponevano nella progettazione del *corpus*: la scelta da operare fra dimensione sincronica e diacronica. La selezione dei testi doveva avere luogo a livello sincronico per consentire, tramite generalizzazione, una descrizione dell'italiano ricorrente nell'uso comune.

[2] Nella fase attuale di sviluppo della disciplina si pone con evidenza il problema della standardizzazione e omogeneizzazione terminologica. Le denominazioni indicate risultano ricorrenti nella maggioranza degli studi.

[3] Nella definizione di Sinclair un *corpus* di riferimento è posto come "designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language and the characteristic vocabulary..." (1995, 29).

Maggiori problemi poneva la scelta fra lingua scritta e lingua parlata. Considerate varie opzioni, pur negli evidenti vantaggi presentati da un *corpus* costituito sia da testi parlati che da testi scritti, si è deciso di procedere, in questa fase della ricerca, dando la preferenza ai testi scritti. La decisione è stata presa sulla base di criteri esterni ed interni. In primo luogo è stata determinata dal panorama linguistico italiano e dalla collocazione che il *corpus* sarebbe venuto ad assumere affiancandosi a opere quali il *Lessico di frequenza dell'italiano parlato* (*LIP,* 1993), il *Lessico di frequenza della lingua italiana contemporanea* (*LIF,*1972), il *Vocabolario elettronico della lingua italiana. Il vocabolario del 2000 (VELI,* 1989) e la *Letteratura Italiana Zanichelli in CD-Rom* (*LIZ* 1993[1], 1995[2] e 1997[3]) per indicare le più significative. In secondo luogo, si è ritenuto preferibile, considerando le trasformazioni che le nuove tecnologie stanno operando nelle modalità comunicative, non porre il problema dei rapporti fra la lingua tradizionalmente indicata come parlato canonico e le estensioni tecnologiche che di questa si realizzano attraverso il mezzo telefonico, radiofonico, televisivo e/o informatico[4].

Si è quindi scelto di costruire un *corpus* sincronico di lingua scritta, i cui testi costitutivi si collocano, pur con qualche approssimazione, in un periodo configurato negli anni '80 e '90[5], con un arco cronologico più ampio per la narrativa, ed appartengono all'italiano che, nei termini posti da Nencioni (1983), può essere definito scritto-scritto[6].

2.2. DIMENSIONE

Maggiore considerazione ha richiesto la definizione della dimensione di un *corpus* che potesse definirsi rappresentativo.

[4] Come si è rilevato (Rossini Favretti 1998a, 2000) la divisione tradizionalmente operata fra scritto e parlato appare inadeguata a rappresentare le varietà linguistiche presenti nella società attuale. Il parlato-telefonico, il parlato-radiofonico ed il parlato-televisivo si articolano, si è visto, in una pluralità di rapporti situazionali in cui si intrecciano modalità discorsive diverse.

[5] Con un'ampiezza temporale richiesta dalla permanenza e dalla "lunga vita" dei libri rispetto ai giornali, per i quali ci siamo limitati alle annate più recenti.

[6] In attesa di approfondire l'analisi dei testi al momento attuale maggiormente ricorrenti in forma elettronica si è convenuto di inserire una sezione opportunamente marcata.

Ad un esame dei *corpora* attualmente disponibili è emerso con chiarezza come non si potesse fare riferimento ad una dimensione standardizzata. Lo sviluppo rapido ed esteso che ha caratterizzato, specie negli ultimi anni, sia l'accessibilità a basso costo dell'*hardware* sia la produzione di programmi *software* sempre più efficienti e di facile utilizzo, ha profondamente mutato i criteri sottesi alla costituzione dei *corpora* più recenti rispetto a quelli di prima[7] o seconda generazione. Se le scelte sottese ai *corpora* di prima generazione, come il Brown Corpus, potevano essere state determinate prioritariamente dalla potenzialità delle tecnologie informatiche, le tecnologie attuali non pongono limiti alle scelte dello studioso, che può estendere la dimensione di un *corpus* fino ad includere le varietà considerate rilevanti ai fini dell'analisi e, all'interno di queste, operare un'adeguata selezione dei testi rappresentativi. Gli sviluppi della tecnologia informatica che si sono avuti negli ultimi decenni, l'attuale velocità nell'elaborazione del materiale ed il basso costo delle unità di memorizzazione consentono oggi di porre il traguardo oltre le otto cifre, offrendo la possibilità di costruire *corpora* di centinaia di milioni di parole come il British National Corpus e la Bank of English. Sembra di potere affermare che, particolarmente per quanto concerne la lingua scritta, lo standard di 1 milione di parole sia ormai sostituito da uno standard di 100 milioni. Ogni generalizzazione, tuttavia, appare controvertibile così come la definizione di un traguardo obbligato. Il Brown Corpus (1967), con 1 milione di parole, 500 campioni di testi scritti, di 2000 parole ciascuno, rappresentativi di generi omogeneamente rappresentati, è ancora considerato da molti studiosi un valido modello. Ed uno dei *corpora* di lingua inglese di più recente costituzione, il Longman Spoken and Written English Corpus – LSWE Corpus – che vede la collaborazione di studiosi come Biber, Johansson, Leech, Conrad e Finegan – presenta una dimensione di circa 40.000.000 parole e contiene 37.244 testi. Testi, si afferma, che variano nella loro lunghezza a seconda del registro.

Un ulteriore aspetto da tenere in considerazione nella definizione del *corpus* è dato dall'introduzione dei *corpora* di monitoraggio. Questi prevedono un costante aggiornamento tramite un flusso di inserimento[8] determinato da un periodico inserimento di dati realizzato da un insieme di filtri, sulla base di una selezione operata sia sui nuovi dati sia su quelli già inseriti[9].

[7] I *corpora* di prima generazione come il Brown e il LOB, basati su un alto numero di brevi estratti, sono oggi generalmente descritti come *corpora* su campioni – *samples corpora* – per distinguerli dai *corpora* su testi completi – *full-text corpora.*

[8] Come si è rilevato, "The 'monitor corpus' [...] is envisaged to be of no finite size but a stream of language in motion, analysed through filters, in real time." (Svartvik 1992, 11).

[9] Da tale prospettiva particolare interesse presenta la Bank of English, un *corpus* della lingua inglese di oltre 300 milioni di parole, configurato nel 1991 come *corpus* di monitoraggio. Il progetto prevede l'inserimento di circa 5 milioni di parole al mese. È da rilevare che nei primi progetti era stato previsto il mantenimento di una dimensione costante, tale da potere essere gestita dai programmi di *software*. "A certain proportion of the data will be stored at any one time, but the bulk will necessarily be discarded after processing. The object will be to 'monitor' such data, from various points of view, in order to record facts about the changing nature of language" (Sinclair 1987, 21). La scelta suscitò le perplessità di alcuni studiosi in particolare per quanto concerneva lo scarto dei testi già inseriti.

La configurazione che il *corpus* di monitoraggio viene ad assumere fa sì che nella definizione della dimensione di un *corpus* cadano quegli aspetti di determinatezza e di permanenza che sono stati caratterizzanti negli ultimi decenni[10]. Il *corpus* presenta una configurazione dinamica che appare tanto più vantaggiosa e rilevante considerando che, con le nuove possibilità date dallo sviluppo dei supporti informatici e delle memorie, al momento attuale non occorre più procedere all'operazione di selezione e di scarto dei testi già inseriti. Appare possibile gestire allo stesso tempo un *corpus* definito nelle sue componenti principali e un *corpus* di monitoraggio, aperto, in grado di registrare le innovazioni e le modifiche ricorrenti nell'uso. La combinazione consente di potere accedere ad un *corpus* disponibile in una forma finita – sia questa data in rete o CD-Rom – suscettibile degli aggiornamenti forniti dal monitoraggio così come dell'introduzione di sottocorpora supplementari rappresentativi di ulteriori varietà.

Si è ritenuto quindi di potere procedere alla progettazione di un *corpus* la cui dimensione, pur essendo configurata come "ampia"[11], non è stata predeterminata ma posta in relazione alla selezione delle varietà linguistiche considerate rappresentative e, in quanto tale, collocata come obiettivo di una fase intermedia della ricerca, successiva alla compilazione di un *corpus* pilota.

2.3. RAPPRESENTATIVITÀ

La definizione della rappresentatività costituisce un momento cruciale nella costruzione di un *corpus*, ma risulta uno degli aspetti maggiormente contro-

[10] Secondo la descrizione di Atkins *et al.* (1992, 5), esso si pone come un *corpus* non archiviato permanentemente, ma i cui testi vengono scannerizzati su una base continua e vengono sottoposti a filtraggio per estrarre i dati costitutivi della banca dati.

[11] L'"ampiezza" di un *corpus* è posta da studiosi come Sinclair come tratto caratterizzante del *corpus*. "A corpus is assumed to contain a large number of words. The whole point of assembling a corpus is to gather data in quantity. The size of corpora continues to increase rapidly, and it would not be sensible to recommend any set of figures." (1995, 21).

versi fra gli specialisti, in particolare per l'ambiguità che si riscontra nell'uso a causa dell'intrecciarsi della connotazione quantitativa e qualitativa.

Se per alcuni studiosi l'estensione dei *corpora* a centinaia di milioni di parole può compensare una scarsa differenziazione delle varietà rappresentate[12], per altri un'ampia differenziazione delle varietà è posta come condizione essenziale di ogni operazione di generalizzazione[13].

Per quanto ci concerne, già nelle prime fasi del lavoro abbiamo ritenuto che il problema della rappresentatività non cadesse con le possibilità di ampliamento del *corpus*, ma anzi potesse venire da questo enfatizzato.

Nonostante l'estensione della dimensione a centinaia di milioni di parole, ogni *corpus* rappresenta un campione limitato della lingua in uso. Un'operazione di campionamento, per quanto estesa, risulta inevitabilmente semplificata rispetto alla complessità del fenomeno in esame. Pur incorporando selezioni probabilistiche nella costruzione del *corpus*, ci è apparso che nel passaggio dal campione alla generalizzazione fosse opportuno prevedere un'approssimazione per gradi che consentisse il massimo di flessibilità e di dinamicità al modello proposto.

Date le difficoltà, vorrei dire di ordine epistemologico, riscontrate nella progettazione di un *corpus* che potesse incontestabilmente definirsi rappresentativo di una lingua o di uno stato di una lingua[14] si è ritenuto di procedere riconoscendo i limiti insiti nella progettazione stessa ed identificando parametri che potessero giungere a controbilanciare quei limiti.

In questo senso appare indicativa l'affermazione di Leech secondo cui può essere posto come parametro della rappresentatività di un *corpus* la possibilità che questo presenta di generalizzare i propri risultati fino ad includere un più vasto *corpus* ipotetico. A tal fine, come si descriverà al punto 3, si sono definiti alcuni criteri di identificazione dei parametri di riferimento che consentissero la costituzione di un insieme di sottocorpora in cui fossero incluse, rappresentate ed adeguatamente bilanciate le principali varietà dell'italiano scritto e, allo stesso tempo, si è configurata la possibilità di giungere all'elaborazione di un modello di costruzione dinamico e adattivo, tale da rispondere alle esigenze ed alle ipotesi di lavoro dei diversi studiosi senza venire meno ai criteri costitutivi del *corpus*.

[12] Si veda, ad esempio, la posizione di ACL Data Collection Initiative (DCI) e del Linguistic Data Consortium, principalmente impegnati nella rapidità della raccolta e della distribuzione di dati.

[13] Esemplificativo in tal senso è il giudizio di Biber che afferma, "analyses must be based on a diversified corpus representing a wide range of registers in order to be appropriately generalized to the language as a whole" (1993b, 180).

[14] Nella presentazione del LSWE Corpus si legge, "No corpus provides a perfect representation of a language, and the LSWE is no exception to this rule" (Biber 1999, 27).

3. Elaborazione del modello di costruzione e strutturazione del *corpus*

3.1. IDENTIFICAZIONE DELLA POPOLAZIONE

Lo spazio impegnato nella descrizione della fase progettuale non mi consente una descrizione analitica delle fasi successive. Mi limiterò ad esporre i passi fondamentali, in primo luogo l'identificazione della popolazione e la definizione dei parametri di selezione. Procederò nella descrizione del progetto considerando in parallelo i punti 2 e 3 dello schema presentato a pag. 40.

Nella letteratura si è posto come tratto definitorio di un *corpus* che i testi costitutivi siano autentici e ricorrenti nella comunicazione sociale. Non si è specificato se questi siano da inserirsi nella loro interezza, o in frammenti che si possano definire rappresentativi. Si tratta, a mio avviso, di un'opzione di primaria importanza, che nella progettazione ha costituito l'oggetto di approfondite riflessioni.

Come si è visto, nei primi *corpora*, quali il Brown, si è operata una standardizzazione dei campioni. L'uniformità dimensionale dei testi è posta come principio costitutivo[15]. Se disaccordo vi è stato, questo si è incentrato sulla dimensione dei campioni. Nell'elaborazione del modello di costruzione, si è ritenuto che, date le condizioni attualmente create dai programmi *software*, il problema non sia dato dalla definizione della dimensione del campione, ma, piuttosto, dalla scelta che deve essere operata fra testi e frammenti di testi. La prima porta inevitabilmente alla mancanza di standardizzazione dei campioni testuali. Si dà raramente il caso che più testi, siano essi giornalistici, narrativi o scientifici, contengano lo stesso numero di parole[16]. La seconda, d'altro lato, comporta, a mio avviso, una più forte presenza della soggettività del ricercatore ed implica una decontestualizzazione delle sequenze selezionate che potrebbe portare, nell'ampia dimensione prevista, ad invalidare la rappresentatività stessa del *corpus*. Si è quindi proceduto privilegiando l'inserimento dei testi nella loro totalità, rispetto alla standardizzazione della dimensione dei campioni.

In un momento successivo, si è proceduto alla definizione delle varietà linguistiche costitutive del *corpus* visto come una collezione di documenti identificabili per caratteri esterni ed interni costituenti un *continuum*, in cui la singo-

[15] Alcuni studiosi, come Biber, affermano che per quanto concerne i tratti grammaticali i risultati sono sufficientemente stabili con campioni testuali di circa 1.000 parole e considerano, quindi, adeguati la maggior parte dei *corpora* esistenti. Altri studiosi, come la Oostdijk, pongono 20.000 parole (*running words*) come lunghezza media dei campioni, in particolare negli studi della variazione linguistica.

[16] Appare una scelta fortemente limitativa porre come criterio discriminante della selezione la ricorrenza nei testi dello stesso numero di parole.

larità della varietà viene a sfumare rispetto alla massa dei dati. Questo costituisce, a mio avviso, un punto importante. Pur inserendo nel *corpus* aree specialistiche, quali il linguaggio burocratico-amministrativo, giuridico, scientifico, si è cercato di fare confluire non una raccolta di testi specialistici ma una varietà di tipologie che si collocano secondo la nostra indagine su un *continuum*, sovrapponendosi ed integrandosi.

3.2 DEFINIZIONE DEI CRITERI DI SELEZIONE E DI COSTRUZIONE

Costante nella letteratura appare il riferimento a criteri esterni ed interni e costante appare l'affermazione che i criteri esterni devono essere privilegiati per ridurre al minimo l'intervento del ricercatore. Si è riconosciuta la validità di queste indicazioni, e, considerando il contesto scientifico in cui questo *corpus* viene a collocarsi così come l'avanzato sviluppo degli studi di *corpus linguistics* che si riscontra a livello internazionale, si è introdotto un ulteriore criterio, "la comparabilità", per non sottovalutare le possibilità che vengono offerte allo studioso dalla comparazione interlinguistica dei *corpora*.

Ai fini della definizione di un primo livello di articolazione del *corpus*, una pregnanza cruciale hanno assunto criteri che definirei di testualità esterna e di comparabilità. Questi hanno portato a configurare un primo livello di articolazione – dato dai sottocorpora – in cui, riducendo al minimo le scelte soggettive del ricercatore, si potesse fare riferimento al alcune macro-varietà identificate sulla base dell'aspetto esteriore o degli elementi materiali dei testi, evidenti nella loro caratterizzazione ed agevolmente comparabili.

Considerata troppo ampia una distinzione che venisse operata fra testi "pubblicati" e "non pubblicati", si è proceduto selezionando le varie forme di pubblicazioni date dalla "stampa", dalla "narrativa"[17], da vari tipi di volumi e di saggi identificabili nella loro varietà come "miscellanea" e sussumendo in una sezione definita "ephemera" i vari testi a mano, a stampa e, principalmente, in formato elettronico caratterizzati dalla loro breve permanenza.

Definite queste macro-varietà si è ritenuto di dovere operare un'ulteriore articolazione – data dalle sezioni ulteriormente scomponibili in sottosezioni – che, ancora basata su parametri esterni, consentisse tuttavia di contestualizzare i dati reperiti. Ad esempio, è apparso chiaro che non si poteva procedere ad un

[17] Non sono stati inclusi testi poetici per il diverso tessuto sintattico-lessicale che questi presentano.Un'analisi della sintattica poetica degli anni '80 mostra che questa, pur avvicinandosi alla lingua comune, standard e substandard, si distacca dalla linearità sintattica corrente. La parola poetica presenta un forte scarto rispetto all'usualità.

campionamento della popolazione "stampa" se non in considerazione di una seconda articolazione, connessa alla realtà socio-culturale nazionale. Questo è stato considerato un momento necessario per giungere a definire, anche se con una certa approssimazione, i componenti della popolazione.

Il riferimento ai parametri descritti al punto 3.1. ha portato a configurare la seguente strutturazione

sottocorpus	STAMPA
sezioni	quotidiana, periodica, supplementi
sottosezioni	nazionale, locale
	specialistica, non specialistica
	connotata, non connotata
sottocorpus	NARRATIVA
sezioni	romanzi, racconti, varia
sottosezioni	italiana, straniera (marcata)[18],
	per adulti, per ragazzi
	poliziesca, di avventure, di fantascienza, delle donne
sottocorpus	PROSA ACCADEMICA
sezioni	volumi, riviste
sottosezioni	scientifica, divulgativa
	scienze umane, naturali, fisiche, sperimentali
	storia, filosofia, arte, critica letteraria, diritto,
	economia, biologia, ecc.
sottocorpus	PROSA GIURIDICO-AMMINISTRATIVA
sezioni	volumi, riviste, documenti
sottosezioni	materiale giuridico (livello normativo, giurisprudenziale e dottrinario), burocratico, amministrativo
sottocorpus	MISCELLANEA
sezioni	volumi, riviste, documenti
sottosezioni	libri religiosi, libri di viaggio, cucina, hobby

[18] Il sottocorpus era inizialmente limitato a testi pubblicati a stampa e/o in formato elettronico, scritti da autori italiani e destinati ad un pubblico di lettori italiani. È stato successivamente allargato ad includere un insieme di libri (di grande tiratura) scritti da autori stranieri per un mercato straniero o internazionale e successivamente tradotti per il pubblico italiano. Questa scelta è stata motivata dall'ampia circolazione che le opere tradotte hanno nel nostro paese e dal peculiare rapporto (quasi di soggezione) che si è determinato fra l'italiano e l'inglese (in particolare l'inglese americano) che si evidenzia non solo a livello lessicale ma anche sintattico.

sottocorpus	*EPHEMERA*
sezioni	lettere, opuscoli, istruzioni
sottosezioni	in formato a stampa, elettronico
	private, pubbliche

Altre varietà potranno essere inserite in una seconda fase del lavoro all'interno di *corpora* supplementari.

3.3. DEFINIZIONE DELLA STRUTTURAZIONE

Definiti i criteri di selezione, si è proceduto alla pianificazione dei sottocorpora, prendendo in primo luogo in esame la dimensione che questi dovevano assumere ed i rapporti che le dimensioni dei vari sottocorpora e delle sezioni dovevano presentare.

In una prima ipotesi si era considerata la possibilità di procedere sulla base di una selezione randomizzata e di correlare la dimensione di ogni sottoinsieme di testi al numero, anche approssimato, dei destinatari di quei testi. Una tale disamina è risultata eccessivamente circoscritta nel privilegiare parametri quantitativi – quali la tiratura e la diffusione – rispetto a parametri qualitativi – quali il tempo e le modalità di utilizzazione dei testi in esame o il livello di attenzione cognitiva. Pur nella difficoltà presentata dall'introduzione di parametri qualitativi – non misurabili – si è ritenuto che il solo dato quantitativo non fosse sufficientemente significativo e che dovesse essere integrato, nella definizione dei rapporti percentuali fra i sottocorpora e le sezioni, da variabili di tipo qualitativo al fine di non sopravvalutare alcune varietà rispetto ad altre. Questa scelta procedurale è stata corroborata da un'analisi di tipo puntuale riferita alle tirature medie del 1997:

GIORNALI[19]	
Quotidiani	2 955 501 360
Settimanali	730 364 544
Mensili	194 607 972
TOTALE	3 880 473 876

LIBRI[20]	
Fiction	119 100 000
Non-fiction	179 400 000
TOTALE	298 500 000

[19] Dati FIEG (1999).
[20] Dati AIE (1999).

Il rapporto 1:12 approssimativamente identificabile fra i testi propri della comunicazione di massa ed i testi del mercato librario non poteva essere accettato come riproducibile nel campione. D'altro lato, esso appariva di tale rilevanza da non potere essere trascurato nemmeno ai fini della comparabilità del *corpus* in costruzione con altri *corpora* quali, ad esempio, il LSWE *corpus* che risulta così costituito come da Figura 1.

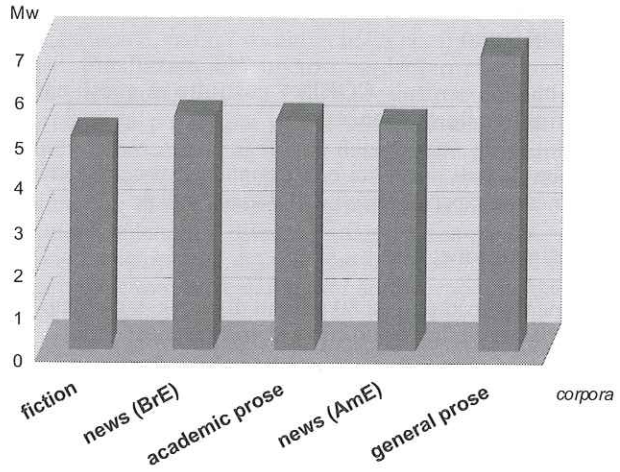

FIGURA 1 – *Articolazione del* corpus *LSWE*

Si è, quindi, convenuto di fissare i rapporti fra le varie fasce di circolazione prendendo, nell'intervallo consentito, il valore che privilegia i testi di circolazione più bassa per non penalizzare alcune varietà quale quella data, ad esempio, dai testi epistolari.

Scelto un ampio insieme di varietà linguistiche, si sono predisposti i documenti per l'inserimento nei singoli sottocorpora e, per aderire ai criteri di rappresentatività, si è proceduto ad una selezione randomizzata dei testi nell'ambito di ogni singolo sottocorpus. Ritenendo di operare in una situazione sufficientemente oggettiva, si è configurata una strutturazione del *corpus* basata sulla seguente articolazione delle macro-varietà precedentemente identificate:

STAMPA	30 milioni di parole
NARRATIVA	20 milioni di parole
PROSA ACCADEMICA	10 milioni di parole
PROSA GIURIDICO-AMMINISTRATIVA	8 milioni di parole
MISCELLANEA	8 milioni di parole
EPHEMERA	4 milioni di parole

3.4. UN *CORPUS* DI ITALIANO SCRITTO – UN MODELLO DEFINITO ED UN MODELLO DINAMICO

Il *corpus* di italiano scritto – CORIS – costruito in questi anni dovrebbe risultare ormai definito nelle sue generalità:

> una raccolta di testi, autentici e ricorrenti nell'uso, in formato elettronico, selezionati come rappresentativi dell'italiano attuale

così come nella dimensione:

> *corpus* generale costituito da 80 milioni di parole aggiornato tramite un *corpus* di monitoraggio inglobato con cadenza biennale.

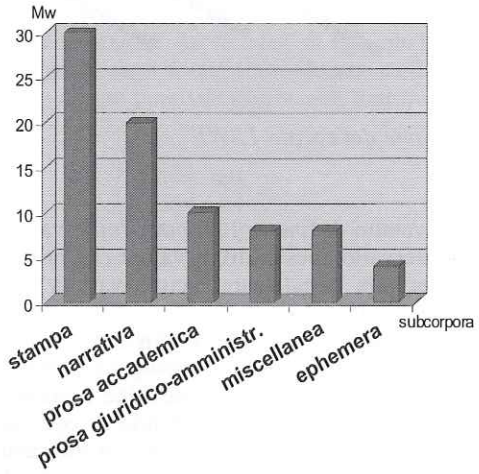

FIGURA 2 – *Articolazione del CORIS*

Il CORIS è stato progettato e costruito come un *corpus* generale di riferimento nell'analisi dell'italiano scritto e sarà messo in linea nel prossimo autunno, presentandosi allo studioso come immediatamente fruibile.

Allo stesso tempo, considerato il ruolo cruciale che viene ad assumere la comparabilità in un *corpus* di riferimento, si è ritenuto fosse opportuno prevedere la possibilità di elaborare una strutturazione alternativa del *corpus* che lo rendesse adattabile alle esigenze dei diversi ricercatori. Accanto al CORIS – Corpus di Riferimento dell'Italiano Scritto – si è configurato il CODIS – Corpus Dinamico dell'Italiano Scritto. Mi dispiace che manchi lo spazio per descriverlo dettagliatamente, ma mi sembra che la configurazione sia già implicita nelle premesse. Finalizzato ad esigenze particolari che possono emergere a livello di analisi interlinguistica, il CODIS presenta una struttura dinamica ed adattiva che fornisce la possibilità di escludere sottocorpora considerati non pertinenti e quindi di procedere, ai fini di determinate procedure di ricerca, ad una selezione di sottocorpora e – aspetto che vorrei sottolineare – delle dimensioni che si ritiene che questi debbano presentare. Il CODIS è predisposto ad essere dinamicamente adattato a diverse situazioni comparative. Tramite procedure sufficientemente semplici è data la possibilità al singolo studioso, qualora questi lo ritenga opportuno, di selezionare i sottocorpora considerati pertinenti e rilevanti scegliendo la dimensione considerata maggiormente funzionale ai fini della ricerca fra le quattro dimensioni indicate o combinandole per costruire dimensioni intermedie (Figura 3).

Sottocorpus	Dimensioni selezionabili (Mw)			
STAMPA	16	8	4	2
NARRATIVA	11	5	3	1
PROSA ACCADEMICA	5	3	1	1
PROSA GIURIDICO-AMMINISTR.	4	2	1	1
MISCELLANEA	4	2	1	1
EPHEMERA	2	1	0,5	0,5

FIGURA 3 – *Articolazione del CODIS*

Un esempio. Data l'articolazione di uno dei *corpora* più recenti, quale il SLWE, che abbiamo appena visto, il ricercatore potrà selezionare nel CODIS le varietà e le dimensioni che lo rendano maggiormente idoneo ad un'analisi interlinguistica e costruirsi un *corpus* così articolato:

STAMPA	4 milioni di parole
NARRATIVA	4 milioni di parole
PROSA ACCADEMICA	4 milioni di parole
PROSA GIURIDICO-AMMINISTRATIVA	2 milioni di parole
MISCELLANEA	2 milioni di parole

Non è che un inizio, ma spero che possa costituire la base per un lavoro comune.

Ringraziamenti

Desidero ringraziare Maria Luisa Altieri Biagi che ha seguito il procedere del lavoro con disponibilità, competenza e consigli.

Bibliografia

Aarts J./ Meijs W. (eds.) (1984), *Corpus Linguistics*, Amsterdam, Rodopi.
AIE (1999), *La produzione libraria italiana del 1997*, Milano.
Aijmer K. / Altenberg B. (eds.), (1991), *English Corpus Linguistics*, London-New York, Longman.
Atkins S. / Clear J. / Ostler N. (1992), "Corpus design criteria", *Literary and Linguistic Computing*, 7: 1, Oxford, Oxford University Press, 1-16.
Baker M. / Francis G. / Tognini-Bonelli E. (eds.) (1993), *Text and Technology: in honour of John Sinclair*, Amsterdam, Benjamins.
Biber D. (1990), "Methodological issues regarding corpus-based analyses of linguistic variation", *Literary and Linguistic Computing*, 5: 257-269.
Biber D. (1993a), "Using register-diversified corpora for general language studies", *Computational Linguistics*, 19: 219-241.
Biber D. (1993b), "Representativeness in corpus design", *Literary and Linguistic Computing*, 8: 4, Oxford, Oxford University Press, 243-257.
Biber D. / Conrad S. / Reppen R. (1998), *Corpus Linguistics. Investigating language structure and use*, Cambridge, Cambridge University Press.
Biber D. / Johansson S. / Leech G. / Conrad S. / Finegan E. (1999), *Longman Grammar of Spoken and Written English*, Longman.
Bortolini U. / Tagliavini C. / Zampolli A. (1972), *Lessico di frequenza della lingua italiana contemporanea*, Milano, IBM Italia.
De Mauro T. / Mancini F. / Vedovelli M. / Voghera M. (1993), *Lessico di frequenza dell'italiano parlato*, Milano, Etas Libri.
FIEG (1999), *La stampa in Italia 1995-1998*, Milano.

Francis W.N. (1982), "Problems of assembling and computerizing large corpora", in Johansson S. (ed.), *Computer Corpora in English Language Research*, Bergen, Norwegian Computing Centre for the Humanities.
Francis W.N. / Kucera H., (1964/1979), *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*, Department of Linguistics, Brown University
Johansson S. (1991), "Times change and so do corpora', in Aijmer K. / Altenberg B. (eds), pp. 305-314.
Johansson S. / Leech G.N. / Goodluck H. (1978), *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*, Department of English, University of Oslo.
Kretzschmar W.A.Jr. / Meyer C.F. / Ingegneri D. (1997), "Uses of Inferential Statistics in Corpus Studies", in Ljung M. (ed.), pp. 167-177.
Kucera H. / Francis N. (1967), *Computational analysis of present-day American English*, Providence, Brown University Press.
Leech G. (1991), "The state of the art in corpus linguistics", in Aijmer K. / Altenberg B. (eds.), p. 27.
Ljung M. (ed.) (1997), *Corpus-based Studies in English, Papers from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17), Stockholm, May 15-19, 1996*, Amsterdam-Atlanta, Rodopi, 1997.
McEnery T. / Wilson A. (1996), *Corpus Linguistics*, Edinburgh, Edinburgh University Press.
Nencioni G. (1983), *Di scritto e di parlato. Discorsi linguistici*, Bologna, Zanichelli.
Oostdijk N. (1988), "A corpus linguistic approach to linguistic variation', *Literary and Linguistic Computing*, 3: 12-25.
Pearson J., (1998), *Terms in Context*, Amsterdam, Benjamins.
Reichard K. / Johnson E.F. (1996), "Using Xforms", *Unix Review*, 84.
Renouf A. (1984), "Corpus development at Birmingham University", in Aarts J. / Meijs W. (eds.).
Rossini Favretti R. (1998a), "Using multilingual parallel corpora for the analysis of legal language: the Bononia Legal Corpus", in Teubert W. / Tognini Bonelli E. / Volz N. (eds.), *Translation Equivalence. Proceedings of the Third European Seminar*, The TELRI Association e.V., Institut für Deutsche Sprache, The Tuscan Word Centre, pp. 57-68.
Rossini Favretti R. (1998b), "Cross-language analysis and large multilingual corpora", *Studi italiani di linguistica teorica e applicata*, XVII: 3, 415-434.
Rossini Favretti R. (1999a), "Scientific discourse: intertextual and intercultural practices", in Rossini Favretti R. / Sandri G. / Scazzieri R. (eds.), *Incommensurability and Translation*, Cheltenham, Edward Elgar, pp. 201-216.

Rossini Favretti R. (1999b), "Equivalenze traduttive in *corpora* giuridici multilingue", *Quaderni di Libri e Riviste d'Italia 43, La traduzione IV*, Roma, Poligrafico e Zecca dello Stato, pp. 47-66.
Rossini Favretti R. (2000), "Oralità, scrittura e variazioni telematiche", *Studi Orientali e Linguistici VII*, (in corso di stampa).
Sinclair J.M. (1986), "First throw away your evidence", in *The English Reference Grammar*, 56-65, Leitner G. (ed.), Tubingen, Niemeyer.
Sinclair J.M. (1987), *Looking up*, London and Glasgow, Collins.
Sinclair J.M. (1991), *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.
Sinclair J.M. (1995), "Corpus typology, a framework for classification", in Melchers G. / Warren B. (eds.), *Studies in Anglistics*, Acta Universitatis Stockholmiensis, LXXXV, Stockholm, Almquist and Wiksell International, 17-34; da "Corpus typology, a framework for discussion", EAGLES document 1-18.
Sinclair J.M. (1996), "Multilingual databases. An international project in multilingual lexicography", in *International Journal of Lexicography*, 9: 3, 179-196.
Sperberg-McQueen C.M. / Burnard, L. (eds.) (1990), *Guidelines for the encoding and interchange of machine readable texts*, TEI, Chicago and Oxford: ACH-ACL- ALLC.
Svartvik J. (ed.) (1992), *Directions in Corpus Linguistics*, Berlin-New York, Mouton de Gruyter.
Teubert W. (1996), "Comparable or parallel corpora?", in *International Journal of Lexicography*, 9: 3, 238-64.
Thomas J. / Short M. (eds.) (1996), *Using Corpora for Language Research*, London-New York, Longman.

ANNOTAZIONE GRAMMATICALE E LEMMATIZZAZIONE DI *CORPORA* IN ITALIANO

FABIO TAMBURINI
Università di Bologna
Centro Interfacoltà di Linguistica Teorica e Applicata "L. Heilmann"

1. Introduzione

Questo lavoro si inserisce nell'ambito del progetto CORIS[1] per la costruzione di un *corpus* di riferimento per l'italiano scritto. Attualmente la realizzazione di un *corpus* di riferimento comporta numerosi problemi che vanno oltre la costituzione del *corpus*. Dimensioni che superano il centinaio di milioni di parole (o *token*) sono divenute ormai lo standard minimale per *corpus* di riferimento, e richiedono quindi opportune metodologie di supporto al lavoro del ricercatore. Effettuare ricerche in *corpora* di queste dimensioni pone problemi, specialmente con lingue ricche di forme flesse come l'italiano: è possibile infatti ottenere, come risultato di una ricerca, una quantità di concordanze largamente superiore a quella gestibile da un essere umano. Esistono numerose tecniche per campionare efficacemente le concordanze ottenute (estrazione casuale, estrazione di una ogni N ecc.), ma non risolvono vari problemi di fondo che possono notevolmente complicare l'analisi dei risultati ottenuti. Supponiamo per esempio che il ricercatore voglia fare uno studio sul verbo 'amare'; le forme flesse del verbo 'amo' e 'ami' corrispondono anche al singolare e al plurale del sostantivo 'amo'. È evidente quindi che senza un meccanismo per la selezione, durante l'impostazione della ricerca, di opportuni criteri restrittivi, ogni calcolo di frequenza sulle concordanze ottenute viene invalidato.

Ci è sembrato quindi utile introdurre nel progetto CORIS opportune metodologie che permettano di specificare parametri aggiuntivi nella ricerca dei termini, per esempio rispetto alle parti del discorso (*part-of-speech tag*) o nei confronti dei lemmi che generano ogni singola parola del *corpus*. Inserire questo tipo di informazioni permette di effettuare ricerche molto più mirate, per esempio rispetto alle forme flesse di un verbo, senza le interferenze generate da

[1] Per la descrizione del progetto si rinvia al contributo di R. Rossini Favretti, in questo volume.