L'introduzione della *corpus linguistics* o *linguistica dei corpora*, nelle università italiane: una ricostruzione 'personale' dagli anni '60 a oggi

MARIA TERESA PRAT ZAGREBELSKY
Università di Torino

1. Premessa

Chi, come me, ha alle spalle un percorso professionale (e umano) ormai piuttosto lungo è tentato, a volte, di guardarsi indietro e tirare le fila di quello che ha fatto. M. A. K. Halliday, un grande protagonista della linguistica del 900, ha premesso al primo volume dei suoi *Collected Works* un suo sguardo all'indietro sulla linguistica di quel secolo intrecciandolo ad alcuni ricordi personali e riferimenti autobiografici (in Webester ed. 2002). Propongo quindi, s*i parva licet*, la mia interpretazione personale di un breve percorso diacronico che va dagli anni '60 ad oggi, che può essere interessante per colleghi della mia generazione e utile ai più giovani per capire un aspetto della storia degli insegnamenti linguistici nelle università italiane.

La mia riflessione riguarderà una tematica particolare, quella della linguistica dei *corpora,* una metodologia che si applica sia alla ricerca linguistica sia alla didattica/apprendimento di una lingua straniera e che è oggi molto diffusa soprattutto per quel che riguarda la lingua inglese. Si tratta di un approccio all'analisi linguistica basato sull'osservazione di insiemi di testi autentici e selezionati in modo da essere rappresentativi del tipo di lingua che si vuole studiare. Dire linguistica dei *corpora* oggi significa parlare di banche di dati linguistici, anche molto grandi, su supporto elettronico e analizzate con l'aiuto di programmi informatici.

Articolerò le tappe della mia ricostruzione per decenni collegando la mia esperienza personale di docente di lingua inglese alle principali tappe dello sviluppo della linguistica dei *corpora* in Europa e nel mondo.

2. Gli anni '60

Chi, come me, si è laureato alla vigilia del 1968, ha frequentato una università di élite dove lo studio delle lingue e letterature straniere signi-

ficava lo studio delle letterature sostenuto da materie storiche e filologiche e da una buona cultura generale. La competenza pratica nella lingua era richiesta, ma affidata prevalentemente alle risorse degli studenti e aveva pochissimi collegamenti con lo studio linguistico nelle sue forme di linguistica storica, di filologia e di glottologia.

In quegli stessi anni il primo *corpus* elettronico di inglese scritto, il *Brown Corpus*, composto da un milione di parole di inglese americano scritto, veniva prodotto nella omonima università negli Stati Uniti e John Sinclair, pioniere dei *corpora* in Gran Bretagna, iniziava a raccogliere a Edimburgo un *corpus* di inglese parlato, trovando poi per i suoi dati una sede tecnologicamente più favorevole nell'università di Birmingham.

3. Gli anni '70

Chi si laureava in quegli anni in lingue e letterature straniere aveva, per un verso, la fortuna di avere un posto assicurato nella scuola e, per altro verso, sperimentava la difficoltà di insegnare la lingua straniera a ragazzi della scuola media inferiore e superiore senza quella preparazione linguistica e metodologica che viene ora fornita, con una riforma che ha tardato a venire, dalle Scuole di Specializzazione per gli Insegnanti della Secondaria (SSIS). Supplivano a queste carenze, nel mio caso, l'entusiasmo per il ruolo di emancipazione sociale della scuola media dell'obbligo e la ricchezza di sperimentazioni di gruppi di insegnanti come il Movimento di Cooperazione Educativa (MCE) e Lingua e Nuova Didattica (LEND).

Alla fine degli anni '70 venni chiamata all'università di Torino – in questo all'avanguardia – come professore incaricato e poi associato di Lingua Inglese per far fronte all'esigenza di organizzare l'apprendimento linguistico. Questa necessità era impellente in una università diventata di massa, dove l'insegnamento delle letterature straniera aveva uno status accademico, ma quello delle lingue straniere era affidato prevalentemente a una figura creata *ad hoc*, quella del lettore di madre lingua straniera. Il professore di lingua inglese, nella mia visione delle cose, doveva tenere corsi di linguistica inglese, cioè di riflessione sistematica e scientifica sulla lingua inglese, e stabilire le finalità e i livelli dell'insegnamento pratico della lingua, impartito appunto da esperti di madre lingua straniera. Ricordo sia la passione che provavo nell'introdurre nell'università italiana una materia fino ad allora esclusa, ma presente all'estero, sia le grandi difficoltà per costruire dal niente una struttura di appoggio composta da libri e riviste, centri linguistici e esperti stranieri

con buona preparazione didattica.

In quegli stessi anni la linguistica dei *corpora* si sviluppava soprattutto in Gran Bretagna e nell'Europa del Nord, passando dalla costruzione di *corpora* 'piccoli' (1 milione di parole) a *corpora* molto grandi (i 100 milioni di parole del *British National Corpus, BNC* e il *corpus* aperto della *Bank of English* di Birmingham, che ha superato ora il mezzo miliardo di parole). Nonostante l'entusiasmo per le potenzialità del computer soprattutto per quanto riguardava risultati di natura quantitativa, la linguistica dei *corpora* non era ancora parte riconosciuta della cosiddetta *mainstream linguistics*, ma restava ai margini, coltivata da piccoli attivissimi gruppi come l'ICAME *(International Computer Archives of Modern and Medieval English)* e diffusa da editori europei come *Rodopi.*

Di questi sviluppi, io, come la maggior parte dei colleghi italiani, non sapevo nulla.

4. Gli anni '80

Nel mio insegnamento universitario sentivo crescere una frattura tra l'apprendimento pratico della lingua, svolto con metodi funzionali-comunicativi dai lettori di madre lingua straniera, e gli obiettivi dei corsi di lingua e storia della lingua inglese. Nei corsi di cui ero responsabile, mi proponevo di presentare i settori e i concetti fondanti dello studio scientifico della lingua (dalla fonologia alla semantica alla pragmatica e testualità) e le diverse teorie linguistiche novecentesche (dallo strutturalismo al generativismo al funzionalismo). In questa panoramica concettuale e storica, svolta in un tempo limitato, fornivo principi generali e alcuni, pochi, esempi prototipici in lingua inglese come "The farmer killed the duckling", "John is eager to please" e "John is easy to please" Cercavo di convincere gli studenti che all'università era necessario riflettere sulla lingua in modo scientifico e che per far questo si dovevano cercare le strutture astratte sotto l'uso linguistico, generale o specialistico che fosse. Tuttavia la scarsezza di dati linguistici autentici che venivano utilizzati nei miei corsi mi dava disagio. Un certo equilibrio era offerto dalla tradizione descrittiva grammaticale e lessicografica inglese, tendenzialmente empirica e eclettica che giungeva in Italia sotto forma di dizionari, grandi grammatiche descrittive e guide all'uso, che venivano consigliati agli studenti come testi di consultazione e riferimento.

Nel 1984 fui colpita da un intervento tenuto da John Sinclair al Simposio londinese che celebrava i 50 anni della fondazione del *British Council.* Sinclair insisteva con forza sulla necessità di lavorare alla descrizione dell'inglese contemporaneo con metodologie informatiche in-

novative e si riferiva all'esperienza di costruzione di grandi *corpora* in corso presso l'università di Birmingham. Proprio in quegli anni cominciarono a essere pubblicati dizionari e grammatiche esplicitamente *corpus-based*. Essi arrivavano in Italia come prodotti già confezionati e un po' misteriosi, pubblicizzati dal ben oliato mercato editoriale britannico.

In Italia, un gruppo di ricerca diretto da Guy Aston (ed. 1988) raccoglieva il *PIXI Corpus*, un *corpus* orale di *service encounters* in inglese e in italiano. Nei convegni seguivo con interesse i risultati delle loro ricerche, ma tutto era ancora per me molto passivo e "cartaceo".

5. Dagli anni '90 a oggi

Il mio ingresso attivo nella linguistica dei *corpora* è avvenuto all'estero prima attraverso periodi di apprendistato presso l'Università di Birmingham e poi con un soggiorno all'Università di Louvain-la-Neuve, in Belgio. In quelle occasioni ho scoperto l'esistenza di diversi tipi di *corpora,* da generali a specializzati, monolingui e plurilingui, commerciali o *ad hoc*, etichettati e no, sincronici e diacronici, di parlanti nativi o di interlingua degli apprendenti. I ricercatori con cui venivo in contatto, che ormai lavoravano da parecchi anni in questo settore, dibattevano sulla natura della *corpus linguistics*, se fosse, cioè, una metodologia applicabile a diversi approcci teorici o se fosse essa stessa un nuovo e rivoluzionario approccio che permetteva di approfondire aspetti del significato delle parole, di scoprire fenomeni come la natura fraseologica della lingua e l'integrazione di lessico e grammatica.

I *corpora* iniziavano a essere utilizzati anche nella didattica sia come fonti di materiali e esercizi per gli autori di testi scolastici e per gli insegnanti sia per l'accesso diretto degli studenti all'osservazione di liste di parole e di concordanze, e, da ultimo, anche sotto forma di produzioni di apprendenti per studiare più sistematicamente le caratteristiche della loro interlingua. Grazie ai contatti con Sylviane Granger e i colleghi di Louvain-la-Neuve riuscii a partecipare a un progetto europeo di raccolta di produzioni scritte di studenti universitari di inglese di diverse nazionalità, il progetto *dell'International Corpus of Learner English,* ICLE (Granger ed. 1998). Diventai così non solo una utente più attiva e consapevole, ma anche la produttrice, in collaborazione con altri, di un *corpus* di saggi argomentativi prodotti da studenti avanzati di inglese di nazionalità italiana.

La *corpus linguistics*, almeno in Europa, non era più ai margini degli studi linguistici, ma era gradualmente diventata un approccio importante

che si opponeva agli approcci generativo-trasformazionali e che si integrava molto bene con approcci funzionali e anche, progressivamente, con l'analisi discorsiva e testuale.

Nel mio contesto universitario la linguistica dei *corpora* sempre di più mi appariva come uno strumento molto utile per conciliare l'esigenza di perfezionamento linguistico pratico, importantissimo nelle lauree triennali da poco entrate in vigore, con quello di osservazione e consapevolezza dei fenomeni che è, e deve restare, fondante nello studio universitario.

La linguistica dei *corpora* tuttavia richiede competenze informatiche e statistiche da parte di chi la insegna e di chi la utilizza, competenze che sono tradizionalmente poco praticate e incentivate in un contesto umanistico. Questo approccio richiede inoltre risorse informatiche, quali laboratori e programmi informatici, oggi ancora poco sviluppate nella università italiana. Nella situazione torinese, una risposta, ancorchè parziale, a queste esigenze viene fornita dal centro linguistico (CLIFU) e non dai dipartimenti (vedi l'elenco dei *corpora* di inglese a tutt'oggi disponibili nell'Appendice 1).

Una volta convintisi dell'importanza della linguistica dei *corpora*, è facile quindi scontrarsi, come a me è successo, con la difficoltà di inserirla nei curricoli degli studenti in maniera concreta ed efficace, cioè in forma attiva e stimolando il loro spirito critico nei confronti dei risultati quantitativi, e la loro sensibilità all' interpretazione 'qualitativa' dei fenomeni linguistici.

A causa di queste difficoltà ho introdotto il concetto di *corpus* nei corsi di Lingua inglese del primo anno inizialmente solo come breve informazione su uno sviluppo importante degli studi linguistici, che è alla base della produzione di nuovi dizionari e grammatiche empiricamente più fondate. Un lavoro più approfondito poteva essere svolto solo in seminari ristretti e nel lavoro di tesi da parte di qualche studente, particolarmente motivato a crearsi gli strumenti necessari.

I miei sforzi di portare la linguistica dei *corpora* a numeri più ampi di studenti sono culminati nell'anno accademico 2004-2005 nella programmazione di un corso di introduzione alla linguistica dei *corpora* indirizzato a un gruppo piuttosto ampio di studenti (circa 125) di un corso avanzato di mediazione linguistica, disposti a frequentare regolarmente e a svolgere lavoro *in itinere*. La programmazione del corso viene riportato interamente qui di seguito, nella versione inglese presentata agli studenti. Il testo mette a fuoco gli obiettivi educativi e professionali, le risorse umane e tecnologiche richieste, i prerequisiti disciplinari e com-

putazionali per gli studenti, l'organizzazione temporale e la strutturazione del sillabo che alterna lezioni frontali e attività di laboratorio e richiede un saggio finale che sviluppi un progetto, anche se limitato, di analisi linguistica attraverso i *corpora*.

TITOLO *Introduction to computer corpus linguistics for students of English: corpora in the description of general English, in EFL teaching/ learning and in the study of specialized languages*

1. Disciplinary and educational aims:

- To introduce advanced students of English to the main principles and applications of computer corpus linguistics, to arouse their curiosity for corpus work and to give them an empowering feeling of autonomy in dealing with a foreign language as the foundation for further professional and academic developments;
- To give students a hands-on experience that is an essential feature of corpus linguistics;
- To merge humanistic and scientific methodologies, and to reinforce the students' familiarity with computing and the Internet;
- To encourage team work and reward regular attendance by involving students in common corpus study projects.

2. Technical equipment and expertise:

- a large classroom with multi-media equipment for general lectures;
- a computer laboratory (25 to 50 places) for guided corpus work;
- a representative selection of the most important corpora available such as the BNC, the ICAME collection, the ICE series, ICLE, and others, with multiple or network licences and computational tools for analysis, e.g. Wordsmith Tools;
- facilities for individual/pair/group work to build 'ad hoc' corpora (e.g. to scan, download and annotate texts);
- a computer expert available during the laboratory sessions and for tutorial guidance in the language centre (CLIFU).

3. Students' prerequisites:

- at least a B2, or better a C1, competence in the use of English;
- knowledge of the main concepts and methods of language study (all students should have attended a foundation course in general linguistics where the language data are mainly taken from their mother tongue, that is Italian);
- knowledge of the main concepts and methods for the study of the English language and the functional analysis of texts;
- basic skills in computing;
- regular attendance of the course.

4. Duration and learning load

The course is worth 10 credits and conventionally would require about 250 hours of study. They will be divided as follows:

– 30 hours of general lectures;
– 30 hours of guided hands-on activities in a computer laboratory, working in pairs or small groups;
– the rest of the time will include individual study, individual practice in the computer laboratory, and the writing of a final corpus-based project.

5. Syllabus

The syllabus will comprise three modules. Each module will have a theoretical and a practical component. There will be weekly between-sessions reading and practical assignments.

1st Module: "Introduction to computer corpus linguistics", 20 hours, five 2-hour general lectures followed by five 2-hour sessions in the computer laboratory.

General topics:
– Corpus linguistics at the beginning of the 21st century;
– The development of computer corpus linguistics: corpus design, size, annotation and number of languages;
– The most important corpora available for English and their characteristics;
– Corpus analysis: corpus-based and corpus-driven approaches.

The main trends and findings of corpus linguistics research: from lexis to discourse:
– Statistics for corpus linguistics (Invited speaker);
– The development of corpus linguistics in Italian and /or other modern languages (Invited speaker/s).

Workshops:
– Some computational concepts and tools for corpus linguistics (Invited speaker);
– Applying Wordsmith tools to a corpus of Italian;
– Quantitative and qualitative analysis of a "first generation" corpus (e.g. Brown/Frown/Lob or Flob);
– Working with a large "second generation" corpus, the British National Corpus, BNC;
– Working with a "tagged" and a "raw" corpus.

2nd Module. "Corpus linguistics and EFL learning and teaching": 16 hours, four 2-hour general lessons followed by four 2-hour sessions in the computer laboratory.

General topics:
– The use of corpora in curriculum design and in the development of reference and teaching materials, e.g. learners' dictionaries and reference grammars;
– The use of corpora in EFL methodology: word lists and concordances;
– Learner corpora with special reference to the 'International Corpus of Learner English , ICLE';
– The use of learner corpora in interlanguage error analysis and Second Language Acquisition (SLA) Research Workshops;
– Corpora as a way to enrich the learning environment;
– Corpus-based and corpus-driven learning through concordances;

– Getting to know "The International Corpus of Learner English, ICLE";
– Practical work on ICLE and on comparable native corpora.

3rd Module: "Working with specialized corpora", 20 hours, five 2-hour general lessons followed by five 2-hour workshops.

General topics:
– Small specialized corpora: design criteria and research findings;
– ESP corpora to study specialized discourse;
– Specialized corpora for terminological work;
– Corpus linguistics and translation: the use of parallel corpora;
– Corpus linguistics and translation: the use of comparable corpora.

Workshops:
– How to build an 'ad hoc' corpus (e.g. to select texts and make them machine readable);
– Software tools and techniques to build specialized glossaries and terminological databases;
– Learning how to align parallel corpora for translation;
– Learning how to build and analyse comparable corpora;
– A corpus-based case study of contrastive rhetoric.

Round-up sessions:
– The pros and cons of computer corpus linguistics;
– How to plan and carry out a small project in computer corpus linguistics.

La realizzazione del corso ha richiesto molto impegno e ha messo a dura prova le limitate risorse umane e computeristiche disponibili soprattutto nel seguire gli studenti nei *workshop* e nella stesura della relazione finale. Tuttavia l'esperienza ha rotto il ghiaccio, per docenti e studenti, e creato le condizioni psicologiche e materiali per considerare i *corpora* come risorse che devono essere note e utilizzabili regolarmente da parte degli studenti così come erano, e continuino ad essere, dizionari, grammatiche e testi di riferimento.

Le ricerche e la letteratura sull'argomento è ormai abbondante non solo in Europa, ma anche in Italia come si può vedere dalla bibliografia selettiva che presenta le opera di alcuni dei più significativi studiosi stranieri accanto a una parte rappresentativa della produzione di studiosi italiani. Alcune università come Forlì-Bologna, Lecce, Padova, Pavia, Pisa, Siena, Torino e Trieste sono molto attive nella ricerca basata su *corpora* e nel loro utilizzo ai fini della didattica linguistica e della traduzione.

A quasi 50 anni dalla sua nascita, attraverso una serie di tappe, si può dire che, anche se con un certo ritardo, la linguistica dei *corpora* é diventata, o sta diventando, *mainstream* anche in Italia.

BIBLIOGRAFIA

ASTON G. (1988), *Negotiating Service. Studies in the discourse of bookshop encounters*, Bologna, CLEUB.

ASTON G. (1997), "Enriching the learning environment: Corpora in ELT", in WICHMANN, A., FLIGELSTONE S., MCENERY T. (1997), *Teaching and Language Corpora*, London, Longman, 51-64.

ASTON G. (1998), "What corpora for ESP?", in PAVESI M., BERNINI G. eds, *L'apprendimento linguistico nell'università: le lingue speciali*, Roma, Bulzoni, 205-226.

ASTON G. (1999), "Corpus Use and Learning to Translate", *Textus* 12, 289-314.

ASTON G. (2001), (ed.) *Learning with corpora*, Bologna, CLEUB; Houston TX, Athelstan.

ASTON G., BURNARD, L. (1998), *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh: Edinburgh Textbooks in Empirical Linguistics.

ASTON G., BURNARD L. eds. (2001), *Corpora in the description and teaching of English*, Bologna, CLEUB.

BERNARDINI S. (2000), *Competence, capacity and corpora,* Bologna, CLEUB.

BERNARDINI S., ZANETTIN F. eds. (2000), *I corpora nella didattica della traduzione. Corpus use and learning to translate. Atti del Seminario di Studi Internazionale*, Bertinoro, 14-15 novembre 1997, Bologna, CLEUB.

BIBER D., CONRAD S., REPPEN R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge, CUP.

DAMASCELLI A. T., MARTELLI A. (2003), *Corpus linguistics and computational linguistics: an overview with special reference to English*, Torino, CELID (2nd edition).

FACCHINETTI R. (2000), "I programmi di analisi linguistica dei corpora: dalla parte degli studenti", in ROSSINI FAVRETTI ed., 109-120.

LEECH G (1997), "Teaching and language corpora: A convergence", in WICHMANN, FLIGELSTONE, MCENERY eds. (1997), *Teaching and Language Corpora*, London, Longman, 1-23.

MAIR C. (2002), "Empowering Non-Native Speakers: The Hidden Surplus Value of Corpora in Continental English departments", in KETTERMAN B., MARKO G., *Teaching and Learning by Doing Corpus Analysis*. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000, Amsterdam/

New York, NY, Rodopi, (with CD-ROM), 119-130.
MCENERY T., WILSON A. (1996), *Corpus Linguistics*, Edinburgh, Edinburgh University Press.
PARTINGTON A. (1998), *Patterns and meanings: using corpora for English language research and teaching*, Amsterdam, Benjamins.
PARTINGTON A. (2003), *The linguistics of political argument. The spin-doctor and the wolf-pack at the White House,* London and New York, Routledge.
PARTINGTON A. / MORLEY J., HAARMAN L (2004), *Corpora and Discourse*, Bern, Peter Lang.
PRAT ZAGREBELSKY M. T. (2004), *Computer learner corpora. Theoretical issues and empirical case studies of Italian advanced EFL learners' interlanguage*, Alessandria, Edizioni dell'Orso.
ROSSINI FAVRETTI R. ed. (2000), *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento,* Roma, Bulzoni
SINCLAIR J. (2003), *Reading concordances*, London, Longman.
SINCLAIR J. (2004), *How to use corpora in language teaching*, Amsterdam, Benjamins.
SINCLAIR J. (2004), *Trust the Text: Language, Corpus and Discourse*, London and New York, Routledge.
STUBBS M. (1996), *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*, Oxford, Blackwell.
TAYLOR TORSELLO C., BRUNETTI G., PENELLO N. eds. (2001), *Corpora testuali per ricerca, traduzione e apprendimento linguistico*, Padova, Unipress.
TOGNINI BONELLI E. (2001), *Corpus linguistics at work*, Amsterdam, Benjamins.

Appendice

Corpora inglese disponibili al CLIFU (Centro Linguistico Interfacoltà per le Facoltà Umanistiche)

International Corpus of English:

British
East Africa
Singapor
New Zealand

ICAME Collection of English Language Corpora:

Written:
Brown Corpus untagged / tagged
LOB Corpus untagged / tagged
Freiburg-LOB (FLOB)
Freiburg-Brown (Frown)
Kolhapur Corpus (India)
Australian Corpus of English (ACE
Wellington Corpus (New Zealand)
The International Corpus of English - East African component

Spoken:
London Lund Corpus
Lancaster/IBM Spoken English Corpus (SEC)
Corpus of London Teenage Language (COLT)
Wellington Spoken Corpus (New Zealand)
The International Corpus of English - East African component

Historical:
The Helsinki Corpus of English Texts: Diachronic Part
The Helsinki Corpus of Older Scots
Corpus of Early English Correspondence, sampler
The Newdigate Newsletters
Lampeter Corpus
Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET)

Parsed:
Polytechnic of Wales Corpus
Lancaster Parsed Corpus (LOB)

British National Corpus (BNC)

International Corpus of Learner English (ICLE) and LOCNESS PIXI