Corpora e interpretazione simultanea

Claudio Bendazzoli

ASTERISCO EDIZIONI

Prefazione

In una società caratterizzata da intensi e sofisticati scambi internazionali all'interprete, quale mediatore linguistico-culturale, è richiesta una professionalità sempre più mirata e qualificata. Al fine di approfondire lo studio dell'interpretazione nelle sue più diffuse modalità e di affinare ulteriormente la didattica di questa disciplina accademica, appare opportuno introdurre la prospettiva della direzionalità vista nel quadro dei rapporti tra coppie di lingue (lingua di partenza-lingua d'arrivo) e con riferimento alla fondamentale questione delle strategie cognitive, linguistiche e testuali attivate dall'interprete. In particolare è opportuno accertare regolarità e caratteristiche di comportamenti traduttivi orali, e quindi comunicativi, in una gamma di combinazioni linguistiche e configurazioni direzionali (da e verso la lingua madre, tra due lingue straniere ovvero tra lingue affini vs. non affini), in una molteplicità di situazioni comunicative, incluso la comunicazione specialistica.

Per l'osservazione e l'analisi dei fenomeni è necessario avvalersi di materiali autentici acquisiti nel corso di conferenze o di trasmissioni televisive o di sperimentazioni ad hoc. Tuttavia, il reperimento di tali materiali, cioè di discorsi originali e relativi discorsi interpretati, non è un'operazione semplice, poiché richiede una metodologia efficace per effettuare la registrazione che è subordinata al consenso da parte di una costellazione di agenti coinvolti nell'evento comunicativo mediato dall'interprete: organizzatori, relatori, interpreti e personale tecnico. Inoltre, al fine di poter rilevare tendenze e regolarità significative nel processo e nel prodotto dell'interpretazione è necessario disporre di una vasta quantità di dati orali da trascrivere e analizzare. Entrambi questi fattori, associati alla generale mancanza di competenze avanzate in informatica e in *corpus linguistics* nel mondo della ricerca in interpretazione, stanno alla base del ritardo dell'introduzione dei corpora negli Studi sull'interpretazione, auspicati già sul finire degli anni Novanta da Miriam Shlesinger (1998).

Nel 2004 presso il Dipartimento di Studi Interdisciplinari su Traduzione, Lingue e Culture (SITLeC) dell'Università di Bologna, sede di Forlì, è nata una intensa collaborazione tra interpreti, linguisti computazionali, esperti di linguistica dei corpora e informatici che ha dato vita ad un gruppo di ricerca interdisciplinare, il *Directionality Research Group*, per studiare la direzionalità in interpretazione mediante la creazione di un corpus elettronico, l'*European Parliament Intepreting Corpus* (EPIC), a cui il dott. Claudio Bendazzoli, titolare di una delle due borse di studio biennali erogate dalla Scuola Superiore di Studi Umanistici dell'Università di

Bologna per questo progetto, ha fornito un contributo fondamentale nelle diverse fasi di concezione del corpus, sviluppo della metodologia, realizzazione e successive applicazioni. EPIC è stato a livello internazionale una delle prime realizzazioni per lo studio dell'interpretazione basato su corpora (*Corpus-based Interpreting Studies*, CIS), un filone innovativo, promettente ed in continua espansione. La sua complessa e laboriosa concretizzazione, che si evince dai capitoli di quest'opera, non sarebbe stata possibile senza una sinergia di sforzi e di competenze diverse tra borsisti e ricercatori, sia nella fase di progettazione che di realizzazione materiale, quest'ultima totalmente per opera dei borsisti dott. Bendazzoli, dott.ssa Sandrelli e dell'assegnista di ricerca dott.ssa Monti.

Tali importanti contributi alla ricerca scientifica sull'interpretazione, proprio per l'ingente impegno di risorse umane e materiali che comportano, sono possibili solo se adeguatamente sostenuti finanziariamente e siamo pertanto grati a chi ha riposto fiducia in questo ambizioso progetto conferendogli le indispensabili risorse economiche.

È molto incoraggiante vedere come un così grande, "epico" sforzo non sia rimasto un prodotto di ingegno isolato, ma ha dato vita a un nuovo corpus che, partendo dalla metodologia sviluppata con EPIC, ne costituisce una naturale evoluzione e, per certi versi, un completamento per la scelta dei dati orali in esso contenuti: si è passati dal linguaggio politico-istituzionale (in italiano, inglese e spagnolo) del Parlamento europeo interpretato simultaneamente verso la propria madrelingua in EPIC al linguaggio specialistico della medicina interpretato verso la lingua madre e verso la lingua straniera con DIRSI-C (*Directionality in Simultaneous Intepreting Corpus*). Nei capitoli dedicati a quest'ultimo, frutto del percorso di dottorato, si potrà apprezzare come il dott. Bendazzoli, in virtù di competenze e metodologie ormai consolidate, sia riuscito a raggiungere traguardi ancora più arditi realizzando un nuovo corpus che oltre all'annotazione, indicizzazione e lemmatizzazione, è caratterizzato da un allineamento testo-audio dei discorsi originali e interpretati, una configurazione questa ottimale per svolgere ricerche mirate sui processi cognitivi e i procedimenti linguistici durante un'interpretazione simultanea, ma che ancora costituisce un raro e pertanto vieppiù prezioso esempio a disposizione della comunità scientifica.

Mariachiara Russo
Presidente del Corso di Laurea magistrale in Interpretazione
Scuola Superiore di Lingue Moderne per Interpreti e Traduttori

Abbreviazioni

CIS	Corpus-based Interpreting Studies
CTS	Corpus-based Translation Studies
CWB	Corpus Work Bench
DIRSI-C	Directionality in Simultaneous Interpreting Corpus
EPIC	European Parliament Interpreting Corpus
LA	Lingua di arrivo
Lingua A	Lingua madre dell'interprete o equivalente
Lingua B	Lingua "attiva", da e verso cui l'interprete sa tradurre
Lingua C	Lingua "passiva" da cui l'interprete sa tradurre verso la lingua A e, eventualmente, verso la lingua B
LLI-UAM	Laboratorio de Lingüística Informática - Universidad Autónoma de Madrid
LP	Lingua di partenza
PE	Parlamento europeo
TA	Testo di arrivo (interpretato)
TP	Testo di partenza (originale)

Introduzione

L'idea di applicare la linguistica computazionale agli studi sulla Traduzione risale all'inizio degli anni Novanta (Baker 1993), quando appare sempre più evidente che «The availability of large corpora of both original and translated text, together with the development of a corpus-driven methodology will enable scholars to uncover the nature of translated texts as a mediated communicative event» (*ibid*., p. 243).

A partire da quel momento, ebbe inizio quella che potrebbe essere considerata una "nuova stagione" per gli studi descrittivi sulla Traduzione (Baker 1995), per diversi motivi. Innanzitutto, una tale metodologia presuppone che i campioni studiati si basino su quantità considerevoli di materiali (soprattutto autentici, ma potenzialmente anche sperimentali) e non su esempi singoli di determinati fenomeni, magari risultato di prove condotte in situazioni realistiche o, addirittura, frutto di introspezione. Inoltre, la ricerca di tendenze significative, nonché di occorrenze molto frequenti di particolari fenomeni, può essere finalmente svolta non solo a fini prescrittivi, ma anche con l'obiettivo di estrapolare "comportamenti traduttivi" da collegare a variabili di natura socioculturale, pragmatica, psicologica e così via. Infine, questa nuova metodologia esige un rigore e una sistematicità tali da consentire di replicare un determinato studio e confrontare così diversi risultati (Laviosa 2004, p. 8). In definitiva, al tradizionale confronto tra testi di partenza (TP) e testi di arrivo (TA) si aggiunge la possibilità di avvicinarsi alla traduzione, scritta e orale, «as a variety of language behaviour that merits attention in its own right» (Baker 1996, p. 175).

Anno dopo anno, i progetti di ricerca dedicati alla realizzazione o all'impiego di corpora in ambito traduttologico sono aumentati costantemente (Olohan 2004, Anderman & Rogers 2008). Questo è vero non solo per quanto riguarda lo studio dei testi originali e dei testi tradotti in quanto "prodotto finito" di attività comunicative (si vedano, fra gli altri, Baker 1999; Kenny 2001; Kruger 2004; Dayrell 2005, 2007; Johansson 2007; Munday 2008; Beeby et al. 2009), ma anche per quanto riguarda la formazione dei traduttori e il potenziale utilizzo dei corpora come strumenti di ausilio alla traduzione (si vedano, tra gli altri, Bowker 1998, Zanettin 1998, Aston 2001, Bernardini 2000, Zanettin et al.

2003, Aston et al. 2004, Bernardini & Castagnoli 2008, Baroni et al. 2006). Pertanto, come ha ampiamente dimostrato Laviosa (2004, pp. 9-17), non è stata smentita la previsione secondo cui all'alba del nuovo millennio «CTS [Corpus-based Translation Studies] was no longer a desideratum, or a research programme, it was a reality. It was here to stay and become a driving force in the discipline for the years to come» (*ibid.*, p. 17). In effetti, questo era già evidente qualche anno addietro, come testimoniato da un'altra raccolta di contributi, pubblicati nel numero 43/4 della rivista *Meta*. La stessa Laviosa (1998a), curatrice di quel numero della rivista, parla del *corpus-based approach* come di un nuovo paradigma di ricerca nei *Translation Studies*. Tuttavia, tra tutti gli altri 14 contributi, solamente Shlesinger (1998) si occupa dell'applicazione di tale approccio o paradigma allo studio della traduzione della comunicazione parlata, ovvero all'interpretazione. È questo un segnale inequivocabile del fatto che lo sviluppo dei corpora elettronici di interpretazione è stato fin dalle prime fasi nettamente rallentato rispetto alla crescita dei corpora di traduzione (scritta).[1] A ben vedere, tale divario si è mantenuto piuttosto costante nel tempo; basti considerare che sempre all'alba del nuovo millennio la riflessione sulla metodologia per creare e analizzare gli *interpretation corpora* aveva forse formulato più interrogativi che risposte (Setton 2002).

Nella descrizione delle possibili tipologie di corpora presentate all'inizio del primo capitolo (§1.2) non sorprende che emergano numerosi riferimenti diretti al campo della Traduzione, poiché il mondo dei testi scritti è anche popolato da innumerevoli esempi di testi tradotti. Appare evidente come la natura statica del linguaggio scritto abbia favorito fin da subito lo sviluppo dei *written corpora* rispetto agli *spoken corpora*, per i quali la raccolta e la trascrizione dei dati hanno invece sempre comportato maggiori difficoltà. Se questo è vero per i corpora orali in generale, lo è ancor di più quando alla lingua parlata è associata l'attività di trasposizione interlinguistica, considerando cioè i corpora di interpretazione. All'interno della categoria dei *Translation corpora* si può pertanto individuare l'esistenza di due filoni di ricerca affini, ovvero i *Corpus-based Translation Studies* (CTS) e i *Corpus-based Interpreting Studies* (CIS), a seconda che ci si occupi di comunicazione e traduzione in lingua scritta o parlata.

L'uso di corpora elettronici negli studi sull'interpretazione è un fenomeno piuttosto recente. Nonostante siano molti i lavori di ricerca che si basano su un "corpus" di materiali analizzati, lo stesso termine è stato spesso impiegato in senso generale e non con riferimento specifico alla linguistica computazionale.
Buona parte delle difficoltà che hanno rallentato l'evoluzione dei CIS sono da imputare non solo alla natura orale dei dati oggetto di studio, ma anche alle fonti

[1] Per una panoramica ampia dei corpora di traduzione cfr. Kenny (2001, pp. 48-72) e Laviosa (2002).

da cui essi possono essere ottenuti e dalla creazione di strumenti di analisi adeguati:

> Research based on spoken corpora is scarcer than that based on written corpora because the former are difficult to compile and work with for a number of reasons, such as the ardous job of data collection, the time consuming and complex transcriptions, and the design of tools which should cater for the idiosyncrasies of the recorded and transcribed material. Despite these obstacles, there is a growing number of researchers engaged in the analysis of spoken corpora, both to reveal specific features of spoken language and to derive pedagogical applications from the results of such analysis.
> (Campoi & Luzón 2007, p. 3)

Le due autrici sopracitate presentano una disamina di diversi progetti basati su *machine-readable spoken corpora*, evidenziandone le potenzialità di ricerca e didattiche in diversi campi, quali l'uso della lingua inglese per scopi accademici e professionali; l'apprendimento linguistico; l'interpretazione. Tuttavia, nessuno dei contributi da loro raccolti tratta specificatamente di alcuna modalità o tipo di interpretazione. Al di là di due riferimenti solo accennati ad alcuni esempi di ricerca attinenti ai CIS, sono più che altro ribadite alcune delle maggiori criticità che si troverebbe ad affrontare chiunque volesse cimentarsi nella creazione di un corpus di interpretazione, sottolineando nuovamente come «the use of corpora offers numerous advantages in the area of interpreting research» (*ibid*., p. 21).

In effetti, è solamente dopo diversi anni dalla formulazione di questo nuovo paradigma di ricerca (Laviosa 1998, Shlesinger 1998) che si comincia ad assistere alla nascita e allo sfruttamento di simili risorse linguistiche anche nel campo dell'interpretazione (Corpas Pastor 2008, pp. 95-98). In tal senso, con il presente lavoro ci auguriamo di contribuire a colmare questa lacuna, per lo meno a livello metodologico. Oltretutto, la disponibilità di nuove risorse linguistiche nel campo dell'interpretazione potrà certamente trarre vantaggio dalle riflessioni già formulate in altri ambiti disciplinari a favore dell'uso dei corpora per attività di ricerca e didattica.

Nello specifico, presentiamo in questa opera due risorse che sono il risultato di due progetti di ricerca nell'ambito dei CIS: EPIC (*European Parliament Interpreting Corpus*) e DIRSI-C (*Directionality in Simultaneous Interpreting Corpus*). Il primo progetto è stato attivato nel 2004 e ha visto la partecipazione di un gruppo di ricerca multidisciplinare (composto non solo da esperti di traduzione e interpretazione, ma anche da linguisti computazionali e informatici), con il coinvolgimento a tempo pieno di un'assegnista di ricerca per il primo anno, assieme a due borsisti di ricerca per il primo biennio. Il progetto è

stato finanziato dal Dipartimento SITLeC e dalla Scuola Superiore di Studi Umanistici dell'Università di Bologna. Il secondo progetto può essere considerato una naturale prosecuzione del primo, ed è stato svolto nell'ambito del programma di Dottorato in Lingue, Culture e Comunicazione Interculturale presso il Dipartimento SITLeC (XXI ciclo). In entrambi gli studi ci si è concentrati sull'interpretazione simultanea fornita da interpreti professionisti in situazioni di lavoro reali, ovvero le sedute plenarie del Parlamento europeo e i convegni internazionali (specialmente di ambito medico e sociosanitario) organizzati nel mercato privato italiano rispettivamente. In EPIC sono presenti tre lingue (italiano, inglese e spagnolo), mentre in DIRSI-C sono incluse due lingue soltanto (italiano e inglese). A ciascuno dei due corpora è associato un archivio multimediale, contenente tutte le registrazioni audio/video dei dati raccolti, di cui solo una parte è stata selezionata per essere inserita nei rispettivi corpora così come sono allo stato attuale.

Oltre a rappresentare i primi esempi di *machine-readable corpora* di interpretazione ad essere stati realizzati in Italia, queste risorse linguistiche racchiudono un potenziale enorme, sia per la ricerca sia per la didattica dell'interpretazione. Gli studi già effettuati e il ventaglio di approfondimenti possibili alimentano un forte entusiasmo nel proseguire questo promettente percorso, con l'obiettivo di contribuire efficacemente all'avanzamento degli *Interpreting Studies* e alla formazione dei futuri interpreti.

Capitolo 1

Teoria e prassi nei C*orpus-based Interpreting Studies*

1.1 Definizione di corpus

> In principle, any collection of more than one text can be called a corpus: the term 'corpus' is simply the Latin for 'body', hence a corpus may be defined as any body of text. It need imply nothing more. But the term 'corpus' when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for.
> (McEnery & Wilson 2001, p. 29)

Nell'aprire questo lavoro in cui saranno discussi i *Corpus-based Translation Studies* (CTS) e, nello specifico, i *Corpus-based Interpreting Studies* (CIS), è parso opportuno offrire una precisazione sul significato del termine "corpus", prima ancora di proporne una definizione. Si vedranno più avanti i motivi di tale scelta che consentirà di mettere in luce una differenza fondamentale all'interno dei CIS.

Come indicato nella citazione sopra riportata, un corpus è qualcosa di più di una semplice raccolta o campionatura di testi, quali espressione concreta di una varietà linguistica (e comunicativa) che si desidera analizzare. Nell'ambito della ricerca linguistica, gli stessi autori sopra citati propongono i seguenti tre elementi come caratteristici di ciò che può definirsi un corpus: rappresentatività, dimensione e formato.

Nel raggruppare una serie di materiali al fine di costituire un corpus, la scelta e la selezione di tali materiali dovrebbero essere basate su un principio di *rappresentatività*. In altre parole, si deve tener conto di criteri con cui poter decidere chiaramente se un testo fa parte o meno della varietà linguistica e comunicativa che intendiamo circoscrivere, per poterlo così inserire o escludere

dal corpus. L'obiettivo ultimo sarebbe ottenere «a sample which is maximally representative of the variety under examination» (McEnery & Wilson 2001, p. 30). Questo obiettivo è evidentemente una sfida notevole, data l'immensità e la varietà dei testi prodotti in qualsiasi ambito comunicativo. Tuttavia, più che chiederci se tale obiettivo sia effettivamente raggiungibile, lo stesso induce a riflettere approfonditamente sulla strategia (criteri) di selezione dei testi da accettare come rappresentativi.

Per quanto riguarda il secondo degli elementi elencato prima, la *dimensione* di un corpus dovrebbe essere definita, soprattutto in funzione della campionatura e della selezione dei testi da includere e che rappresentano la varietà linguistica e testuale oggetto di studio. In realtà, vi è anche la possibilità che un corpus sia "aperto", cioè potenzialmente in continua espansione. Questo tipo è anche conosciuto come *monitor* corpus (Sinclair 1991, pp. 24-26) e non prevede un tetto massimo di raccolta dei materiali.

Infine, stando al terzo e ultimo elemento indicato, il *formato* dei testi inclusi nel corpus dovrebbe essere *elettronico*, di modo che i testi in tale formato possano essere elaborati da un computer e analizzati attraverso specifici programmi informatici di linguistica computazionale. Nel caso dei corpora di lingua parlata, questo è uno dei fattori che presuppongono che i dati registrati siano trascritti (§1.3.3). Tuttavia, un corpus i cui testi o trascrizioni sono disponibili solo su supporto cartaceo (stampati) consente di effettuare unicamente analisi di tipo "manuale"; in casi sempre più frequenti questa opzione ormai non è più contemplata, in quanto assai poco praticabile se si pensa alle grandi dimensioni delle campionature costituite in quasi tutti i progetti di ricerca (McEnery et al. 2006, p. 6).

A questo punto, un corpus può essere così definito: «[…] a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration» (McEnery & Wilson 2001, p. 32).[1]

Un'ulteriore caratteristica rilevante di un corpus (ma non essenziale) è data dalla possibilità di aggiungere all'interno dei testi che lo compongono tutta una serie di informazioni utili all'estrazione dei dati, potendo dunque far sì che questi siano "filtrati" in modo da far emergere particolari aspetti dei testi che costituiscono il corpus. Tale "aggiunta" ed "esplicitazione" di informazioni

[1] Sulla definizione di "corpus" si veda la breve, ma efficace panoramica in McEnery et al. (2006, pp. 4-5): «There are many ways to define a corpus […], but there is increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety» (*ibid*., p. 5). Un interessante aspetto aggiuntivo proposto dagli autori riguarda la possibilità che il corpus sia *balanced* o meno (in quest'ultimo caso si parla di *specialized* corpora, ovvero porzioni limitate, cioè sottocorpora, di un corpus più grande). Nel trattare i linguaggi specialistici, Bowker & Pearson (2002) danno la seguente definizione di corpus, la quale è pienamente compatibile con quelle che abbiamo già riportato: «A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria» (*ibid*., p. 9).

prende il nome di *annotazione* e *codifica*, operazioni che possono essere effettuate seguendo diversi standard. Ad esempio, si possono aggiungere informazioni linguistiche e/o metalinguistiche in forma di etichette (*tags*), applicate a ogni singolo testo nella sua interezza (spesso le informazioni metatestuali sono raccolte in un *header* – una sorta di intestazione in cui sono elencati diversi campi e parametri di classificazione), a una porzione di esso o ad ogni singolo *token*. Pur non essendo incluso tra i fattori "essenziali" nella definizione di un corpus, l'annotazione apporta un valore aggiunto notevole (Leech 1997a), in quanto consente la realizzazione di ricerche mirate, altrimenti rese più ostiche, se non impossibili, dalla diversità strutturale di ogni singolo testo e dalla ricchezza morfologica di qualsiasi lingua. Infine, il corpus dovrebbe essere reso accessibile alla comunità scientifica di modo che possa essere studiato e usato come strumento didattico o di ricerca, a seconda dei materiali in esso raccolti.

1.2 Tipi di corpus

Esistono diversi fattori che contribuiscono a determinare varie tipologie di corpus (Baker 1995, Shlesinger 1998). È possibile, infatti, fare una distinzione a seconda del tipo di lingua (scritta, orale o altro), nonché del numero di lingue coinvolte nei testi selezionati per un corpus. Inoltre, l'impiego di testi originali e di testi tradotti, assieme o separatamente, aumenta ulteriormente le possibilità di realizzazione. Infine, alcune caratteristiche della comunità linguistica di riferimento e del tipo di testo o evento linguistico preso in esame possono conferire al corpus una "identità" dai tratti ancor più particolareggiati (per esempio, a seconda che i testi in una determinata lingua siano prodotti da parlanti madrelingua o da parlanti non nativi). Sulla base dei fattori appena menzionati, si possono elencare le seguenti tipologie di corpora (molte categorie sono associabili a uno stesso corpus e non si autoescludono vicendevolmente):

Tabella 1.1 Tipologie di corpora (1).

Tipologia di corpus	Descrizione
corpus di lingua scritta (*written corpus*)	Contiene testi appartenenti alla comunicazione scritta
corpus di lingua parlata (*spoken corpus*)	Contiene trascrizioni di testi (eventi linguistici) appartenenti alla comunicazione parlata
corpus monolingue (*monolingual corpus*)	Contiene testi nella stessa lingua
corpus bilingue e multilingue (*bilingual and multilingual corpus*)	Contiene testi in due o più lingue diverse tra loro (si avrà quindi un sottocorpus per ciascuna lingua)
corpus parallelo (*parallel corpus*)	Contiene testi originali e le loro traduzioni in un'altra lingua (si avrà quindi un sottocorpus di TP e uno o più sottocorpora di TA). Può essere monodirezionale o bidirezionale, a seconda che le traduzioni siano effettuate soltanto da una lingua X a una lingua Y o anche viceversa, da Y a X
corpus comparabile / paragonabile (*comparable corpus*)	Contiene testi in una stessa lingua, ma provenienti da sottocorpora diversi per tipo di campionatura svolta. Ad esempio, si può avere un sottocorpus di testi narrativi in lingua italiana e un sottocorpus di traduzioni in italiano di testi narrativi scritti originariamente in altre lingue; o ancora si possono avere diversi sottocorpora per diverse varietà della stessa lingua. Alcuni autori (come McEnery et al. 2006) considerano questi ultimi esempi *comparative corpora*, mentre definiscono *comparable corpora* i corpora contenenti testi di lingue diverse (non traduzioni) ma basati su pari criteri di campionatura, rappresentatività e dimensione[2]
corpus multimodale (*multimodal corpus*)	Si propone di descrivere molteplici livelli semiotici che concorrono alla costruzione del significato di qualsiasi entità appartenente al mondo animato e inanimato; può quindi riguardare non solo testi scritti e orali, ma anche film, oggetti, prodotti artistici, ambienti e così via (Baldry & Thibault 2001). A tal fine, si struttura affiancando in parallelo i diversi livelli semiotici presi in esame. Per esempio, in un corpus multimodale di materiale pubblicitario si potrebbero individuare i seguenti livelli: colore, musica, testo, distribuzione nello spazio, ecc. (Baldry & Thibault 2005). Un raro esempio di corpus multimodale di interpretazione è stato realizzato per la modalità di interpretazione in lingua dei segni (Kellet Bidoli 2007)

[2] Per un approfondimento sulla terminologia utilizzata in letteratura per indicare diverse tipologie di corpora si veda Zanettin 2001 (sezioni 3 e ss.) e Ulrych (2001). Diverse configurazioni ottenibili con testi originali e testi tradotti sono illustrate da Johansson (1998).

corpus multimediale (*multimedia corpus*)	Raccoglie contenuti multimediali, comprendenti quindi le tre dimensioni scritto-visivo-sonoro (McEnery & Wilson 1997). Alcuni esempi significativi di corpora multimediali sono il corpus ELISA (Braun 2006a), contenente videointerviste e relative trascrizioni per l'insegnamento/ apprendimento della lingua inglese e lo sviluppo di risorse simili a fini pedagogici; il corpus FORLIXt (Heiss & Soffritti 2008, Valentini 2009) contenente materiale filmico originale e doppiato in più lingue (italiano, francese e tedesco) per lo studio della traduzione audiovisiva; il corpus *Marius* (de Manuel 2003b) con videoregistrazioni di discorsi presso il Parlamento europeo e consessi internazionali, accompagnate dalla trascrizione e classificate secondo diversi gradi di difficoltà per l'apprendimento dell'interpretazione simultanea e consecutiva
corpus intermodale (*intermodalcorpus*)	Appartiene specificatamente all'ambito traduttologico. Contiene più TA di uno stesso TP, prodotti attraverso differenti modalità traduttive come, ad esempio, traduzione scritta, interpretazione simultanea, interpretazione consecutiva, ecc. (Shlesinger 2008, p. 240)

In aggiunta alle tipologie illustrate, i corpora possono essere ulteriormente classificati secondo il tipo di varietà e comunità linguistica rappresentate (McEnery et al. 2006, pp. 59-69; Bowker & Pearson 2002, pp. 11-13):

Tabella 1.2 Tipologie di corpora (2).

corpus generale di riferimento (*general* o *reference corpus*)	Corpus di grandi dimensioni, rappresentativo dell'uso generale di una lingua e comprensivo di una gamma di tipologie testuali di varia natura (di lingua scritta ma anche di lingua parlata). Un esempio classico è il *British National Corpus* (BNC)
corpus specialistico (*specialized corpus*)	Corpus circoscritto a un particolare aspetto o varietà di una lingua, comprensivo di testi tipici di un determinato ambito e tipo
corpus sincronico (*synchronic corpus*)	Corpus contenente materiale rappresentativo di uno stesso periodo di tempo
corpus diacronico (*diachronic corpus*)	Corpus contenente materiale rappresentativo di un lasso di tempo e in grado, quindi, di mostrare l'evoluzione della lingua (Facchinetti & Rissanen 2006)
corpus aperto o monitor (*open corpus* o *monitor corpus*)	Corpus espandibile potenzialmente all'infinito, senza un tetto massimo prestabilito rispetto alla quantità dei materiali in esso raccolti
corpus chiuso (*closed corpus*)	Corpus la cui dimensione massima è definita. Non sono pertanto previste ulteriori aggiunte di materiali dopo che il limite prestabilito è stato raggiunto

learner corpus	Corpus contenente testi provenienti da parlanti non nativi che stanno apprendendo la lingua in cui tali testi sono prodotti. Un esempio per la lingua finlandese è il progetto ICLFI – *International Corpus of Learner Finnish* (Jantunen 2008). Sono diffusi anche in ambito traduttologico per studiare non solo la L2, ma anche l'acquisizione e l'uso delle strategie traduttive da parte di traduttori in formazione (per una rassegna si veda Castagnoli 2008, 2009)
web-based corpora o *web as corpus (WaC) approach*	Corpora i cui materiali sono tratti direttamente da Internet, caratterizzati da un alto grado di rappresentatività (si possono raggiungere dimensioni che superano facilmente il miliardo di *token*) e facili da aggiornare. In base a questo nuovo approccio, la Rete stessa può essere esplorata come un corpus globale attraverso l'uso diretto dei motori di ricerca o l'impiego di programmi appositi che consentono di gestire più autonomamente l'output ottenuto sempre attraverso i motori di ricerca; in alternativa, la Rete può essere utilizzata come una fonte da cui scaricare enormi quantità di materiali (pagine web e documenti), effettuando una preselezione e potendo poi esplorarli autonomamente, senza dipendere quindi dagli algoritmi che sottostanno al funzionamento dei motori di ricerca commerciali (Baroni & Bernardini 2006, Baroni et al. 2009, Ferraresi 2009)

1.2.1 I corpora di lingua parlata

Tra le diverse tipologie di corpora che sono state messe a fuoco, solo alcune di esse assumono particolare rilevanza all'interno del presente studio. Considerando che i corpora di interpretazione sono *in primis* corpora di lingua parlata, vale la pena esaminare alcuni dei principali progetti di ricerca svolti in questo ultimo ambito.[3] In generale, gli esperti concordano nel riconoscere che «The creation of a spoken corpus is not as straightforward as that of a written one […]» (Sinclair 1991, p. 16). Le maggiori difficoltà non risiedono solo nel gravoso compito di trascrizione dei dati orali, ma anche nelle procedure di raccolta e di gestione dei dati stessi, nonché nel tipo di annotazione con cui arricchire le trascrizioni (sfide che, come vedremo nelle sezioni successive, hanno rallentato notevolmente lo sviluppo dei CIS rispetto ai CTS).

[3] Non essendo possibile fornire un elenco esaustivo dei numerosi progetti realizzati ad oggi, specialmente per la lingua inglese, si rimanda per esempio a Edwards (1993b), alla banca dati e alle risorse messe a disposizione dal *Linguistic Data Consortium* (LDC) e alla banca dati della *European Language Resources Association* (ELRA).

Ciononostante, è lecito affermare che i corpora orali non appartenenti ai CIS godono di una tradizione più matura rispetto a quanto è stato possibile realizzare in ambito traduttologico.

Un primo aspetto da puntualizzare è che esistono sia corpora di riferimento in cui è compresa una parte di dati riguardanti la comunicazione parlata (unitamente ai testi in lingua scritta), sia corpora esclusivamente orali (McEnery et al. 2006, pp. 62-64). Esempi notevoli di corpora di riferimento contenenti uno o più sottocorpora di lingua parlata sono il BNC – *British National Corpus* (per la lingua inglese) e il CREA – *Corpus de la Real Accademia* (per la lingua spagnola).[4] Dall'altra parte, tra i corpora esclusivamente orali vi sono progetti di grande portata anche per lingue diverse dall'inglese (probabilmente la lingua che può contare sul maggior numero di risorse di questo tipo),[5] ad esempio lo *Spoken Dutch Corpus* (Goedertier et al. 2000, Oostdijk et al. 2002), CoSIH – *Corpus of Spoken Israeli Hebrew* (Izre'el et al. 2001), e il *Czech Spoken Corpus* (quest'ultimo in realtà facente parte di un progetto più ampio, nel quale sono inclusi ben quattro corpora orali: ORAL2008, ORAL2006, PMK – *Prague spoken corpus* e BMK – *Brno spoken corpus*).[6] Per quanto riguarda le lingue romanze,[7] con il progetto C-ORAL-ROM (Cresti & Moneglia 2005) è stato profuso uno sforzo congiunto tra diverse unità di ricerca al fine di sopperire alla carenza di risorse linguistiche, quali i corpora orali nelle lingue romanze. Tale progetto ha portato alla creazione di ben quattro corpora di lingua parlata in portoghese, francese, spagnolo e italiano. Ciascun corpus contiene circa 300.000 parole, per un totale di 772 testi, 121 ore di registrazione e 1.427 partecipanti, attingendo da situazioni comunicative di vario genere (dalla conversazione spontanea in ambito familiare alle trasmissioni radiotelevisive, ecc.). È interessante notare che coordinando le quattro diverse unità di ricerca è stato possibile, per molti versi, avvicinare i sistemi di trascrizione, codifica e annotazione per le quattro lingue coinvolte. Ad esempio, per tutti i materiali sono state prodotte trascrizioni ortografiche in un

[4] Come riportato da Munday (2008, p. 234 nota 17), «The BNC comprises around 110 million words of naturally-occurring (mainly British) English taken from a range of sources, including fiction and newspapers but also some spoken language and informal written material such as advertisements and fliers. The project ended in 1995 and contains texts predominantly published in the 1980s and 1990s. The RAE current corpus (CREA), contains a similar number of words and range of genres and text types and has a fifty-fifty split of peninsular Spanish and Latin American texts». Altre risorse in lingua spagnola (ma non solo) di particolare interesse sono state create dai ricercatori del *Laboratorio de Lingüística Informática* (LLI) presso la *Universidad Autónoma de Madrid* (UAM), tra cui il corpus CHIEDE (Garrote Salazar 2008) sul linguaggio spontaneo infantile.

[5] Si vedano, per esempio, i diversi contributi raccolti nelle pubblicazioni risultanti dalla serie di conferenze ICAME (*International Computer Archive of Modern and Medieval English*) intitolate *International Conference on English Language Research on Computerized Corpora* (Lee 2008).

[6] Si veda la Sitografia per i riferimenti specifici di ciascuno dei progetti menzionati.

[7] Per una panoramica più ampia di risorse per le lingue romanze si veda Pusch (2002). Un'altra rassegna sintetica sui corpora per l'italiano, le lingue romanze e il panorama anglosassone è condotta da Cresti (2000a, pp. 13-21).

formato testuale standard conforme al formato CHAT (MacWhinney 2000, si veda §1.3.3.2). Esso prevede l'inserimento di dati meta-testuali all'inizio di ogni trascrizione (*header*) e consente di rappresentare in forma scritta l'interazione dialogica. Inoltre, i testi trascritti presentano una suddivisione in enunciati, la cui annotazione è stata effettuata attraverso un metodo euristico, basandosi cioè sui tratti prosodici e sul giudizio percettivo degli stessi da parte dei trascrittori. È inoltre presente un'annotazione prosodica in corrispondenza dei cosiddetti *prosodic breaks*; essi sono segnalati da una barra singola [/] o da una doppia barra [//], a seconda che si tratti di *non-terminal breaks* o *terminal breaks*, rispettivamente. Due ulteriori livelli di annotazione disponibili sono l'annotazione grammaticale (*POS-tagging*) e la lemmatizzazione per ogni singolo *token* (§1.3.4.1). A questo proposito, uno dei risultati di maggior rilevanza di questo progetto è stata la possibilità di testare e valutare la *performance* dei programmi di annotazione automatica su testi "orali" (cioè le trascrizioni). Normalmente, se applicati a testi scritti, questi programmi (*taggers*) funzionano correttamente perché vi è completa consonanza tra la lingua del testo e la grammatica interna o le regole probabilistiche su cui è basata l'assegnazione automatica di ciascuna etichetta (morfologica, grammaticale, e così via, §1.3.4). Per contro, i testi tra-scritti riflettono da vicino le caratteristiche salienti della comunicazione parlata, quali le ripetizioni, le false partenze, le riformulazioni, le parole incomplete e mal pronunciate; si tratta di fenomeni che non sono contemplati nelle regole strutturate su cui si basa il funzionamento dei *taggers* e che potrebbero incidere negativamente sul loro tasso di successo nell'assegnare le etichette corrette. Nonostante il grande sforzo di mantenere un elevato grado di uniformità tra i corpora C-ORAL-ROM, le inevitabili differenze tra le quattro lingue romanze di questo progetto hanno comportato comunque l'adozione di repertori di etichette (*tagset*) specifici[8] per ciascun sottocorpus. «Nontheless, in order to ensure comparability within the whole corpus, a compulsory minimal threshold of information has been established in the tag codes» (Cresti & Moneglia 2005, p. 52).

Il corpus italiano nel progetto C-ORAL-ROM (Cresti et al. 2005) ha attinto dai sottocorpora raggruppati in LABLITA (Cresti 2000a, 2000b; Moneglia 2005): Corpus dell'italiano adulto spontaneo, Corpus della prima acquisizione dell'italiano e Corpus della lingua cinematografica e dei media. Sempre per l'italiano parlato, esistono ulteriori progetti coordinati tra più unità di ricerca (Albano Leoni 2005), quali API (Archivio del Parlato Italiano) e

[8] Una caso particolarmente interessante è dato dall'etichetta MD, utilizzata nei sottocorpora spagnolo e portoghese per indicare i segnali discorsivi. A questo proposito, sono state considerate sia parole individuali, sia le cosiddette *multiwords*, ovvero stringhe di due o più *token* che per le loro funzioni discorsive sono considerate un'unità singola non frammentabile. In spagnolo (Moreno Sandoval et al. 2005, Moreno Sandoval & Guirao 2006) troviamo per esempio la stringa *es decir* annotata come un un'unica entità (*es_decir* MD).

AVIP (Archivio delle varietà dell'Italiano Parlato), focalizzati sul parlato dialogico; oltre a questo formato interazionale, il progetto CLIPS (Corpora e Lessici di Italiano Parlato e Scritto) è anche mirato al parlato radiotelevisivo, telefonico e letto.[9] In modo simile, ma sulla base di diverse campionature di registrazioni, sono stati condotti altri studi che hanno portato alla realizzazione di corpora e risorse conosciute come LIP[10] (Lessico di Frequenza dell'Italiano Parlato, De Mauro et al. 1993), LIR (Lessici dell'Italiano Radiofonico, Alfieri & Stefanelli 2005) e CIT (Corpus di Italiano Televisivo, Spina 2005). Un interessante esempio di ricerca contrastiva basata sui corpora tra la lingua (e la cultura) italiana e inglese è il progetto PIXI (*Pragmatics of Italian/English Cross-Cultural Interaction*, Gavioli & Mansfield 1990), con un corpus in cui sono raccolte 379 conversazioni avvenute all'interno di librerie in Italia e in Inghilterra, registrate e trascritte seguendo convenzioni ispirate al sistema jeffersoniano (§1.3.3).

Spostando la nostra attenzione dalla lingua italiana alla lingua inglese, troviamo un grandissimo numero di progetti di *spoken corpora*,[11] tra cui ve ne sono di specifici per l'inglese britannico, l'inglese americano e l'inglese parlato come *lingua franca* o come L2 da parlanti non nativi. Uno dei corpora di dimensioni maggiori è il CANCODE (*Cambridge and Nottingham Corpus of Discourse in English*), contenente trascrizioni tratte da registrazioni effettuate in varie località del Regno Unito tra il 1995 e il 2000, per un totale di cinque milioni di parole. In esso, quindi, sono rappresentate molteplici comunità e varietà linguistiche dell'inglese britannico parlato in modo spontaneo. Tale risorsa è stata e continua ad essere utilizzata per realizzare opere lessicografiche, grammatiche e studi linguistici di varia natura.

Intorno ai tre milioni di parole, invece, si attesta un corpus di inglese parlato "telefonico" chiamato SWITCHBOARD (Godfrey et al. 1992). Questo corpus raccoglie più di 2.400 conversazioni telefoniche della durata di sei minuti circa, per un totale di oltre 240 ore di registrazioni e più di 500 partecipanti provenienti da diverse zone degli USA. Il particolare canale di trasmissione utilizzato (il telefono) ha consentito di raccogliere una vasta gamma di materiali sul piano diatopico e di svolgere la procedura di raccolta dei dati in maniera semiautomatica. Si tratta a tutti gli effetti di una situazione comunicativa

[9] Per una raccolta di risorse sull'italiano parlato si veda il sito dell'osservatorio Parlare Italiano.

[10] Con il progetto Badip – Banca dati dell'italiano parlato (Schneider 2002) è stata realizzata una versione elettronica e *online* di questo corpus.

[11] Stando a Leech (1997a), «it was not until the mid-1970s that a first major attempt was made to establish a computer corpus of spoken language» (*ibid*., p. 10), arrivando a creare il corpus LLC –*London-Lund Corpus* (Svartvik 1990). Questo corpus è stato creato utilizzando materiali registrati dal 1960 nell'ambito del progetto *Survey of English Usage corpus* (500.000 parole) presso lo *University College London*. Un altro corpus che ha segnato la storia dell'evoluzione degli *spoken corpora* è il SEC – *Lancaster/IBM Spoken English Corpus* (Knowles 1993), di dimensione inferiore rispetto al precedente, ma con diversi livelli di annotazione (grammaticale, sintattica e prosodica).

artificiale, dalla quale non è detto che si possano trarre informazioni sui meccanismi conversazionali spontanei. Ciononostante, il principale obiettivo di questo progetto in realtà non riguardava lo studio della lingua in sé, bensì l'approfondimento di questioni attinenti al riconoscimento vocale automatico (è questa una delle tante applicazioni degli *spoken corpora*).

Un altro esempio particolarmente rilevante per l'inglese americano è il corpus MICASE – *Michigan Corpus of Academic Spoken English* (Simpson et al. 2007). In questo corpus di circa 1,8 milioni di parole sono raccolti vari eventi comunicativi che hanno avuto luogo nell'ambito delle attività universitarie della *University of Michigan*.[12] Per poter realizzare il corpus in formato elettronico, uno dei passi fondamentali è stato la codifica e l'annotazione degli eventi registrati (e trascritti), tra i quali vi sono lezioni frontali e seminari che potrebbero richiamare le dinamiche di una conferenza-convegno. Di seguito riportiamo sinteticamente gli attributi che sono stati definiti in merito agli eventi e ai partecipanti:

Tabella 1.3 Attributi utilizzati per la codifica dei materiali in MICASE.

speech event attributes	classroom events non-class events academic division academic discipline participant level primary discourse mode
speaker attributes	gender age group academic role native speaker status first language

Le voci riportate nella Tabella 1.3 sono a loro volta definite attraverso l'inserimento di ulteriori specifiche, costituite da categoria, codice e definizione/commenti, a seconda delle caratteristiche peculiari di ogni evento e partecipante.[13] Nello specifico, in vista della classificazione dei materiali da includere in un corpus (si veda la tassonomia sviluppata nel progetto DIRSI §4.1, §4.2.3.3), destano particolare interesse le categorie stabilite per gli attributi indicati con *classroom events* e *non-class events*. Essi sono riportati di seguito nella Tabella 1.4 e nella Tabella 1.5 rispettivamente:

[12] Un'altra risorsa sull'inglese americano è il *Santa Barbara Corpus of Spoken American English* (Du Bois et al. 2000, 2003; Du Bois & Englebretson, 2004, 2005). Questo corpus ha dimensioni decisamente inferiori rispetto al MICASE (circa 249.000 parole) e include registrazioni effettuate in situazioni quotidiane o da mezzi di comunicazione. Il materiale è confluito in un altro corpus di riferimento, con dati sia di lingua scritta che di lingua parlata, chiamato ICE – *International Corpus of English* (Greenbaum 1996).

[13] Per maggiori dettagli, si rimanda alla pagina web del progetto MICASE.

Tabella 1.4 Categorie per i *classroom events* nel corpus MICASE.

LARGE LECTURES	Lecture class; class size = more than 40 students
DISCUSSION SECTIONS	Additional section of a lecture class designed for maximum student participation; may also be called recitation
LAB SECTIONS	Lab sections of science and engineering classes; may include problem solving sessions
SEMINARS	Any class defined as a seminar (primarily graduate level)
STUDENT PRESENTATIONS	Class other than a seminar in which one or more students speak in front of the class or lead discussion

Tabella 1.5 Categorie per i *non-class events* nel corpus MICASE.

ADVISING SESSIONS	Interactions between students and academic advisors
COLLOQUIA	Departmental or University-wide lectures, panel discussions, workshops, brown bag lunch talks, etc.
DISSERTATION DEFENSES	Ph.D. theses defences
INTERVIEWS	Interviews for research purposes
MEETINGS	Faculty, staff, student government, research group meetings, not including study group meetings
OFFICE HOURS	Held by faculty or graduate student instructors in connection with a specific class or project
SERVICE ENCOUNTERS	Library, computer center, financial aid office services
STUDY GROUPS	Informal student-led study groups, one time or on-going
TOURS	Campus, library, or museum tours
TUTORIALS	One-on-one discussions between a student and an instructor or peer tutor

Sempre in ambito accademico, ma in un contesto britannico, una risorsa simile a MICASE è il corpus BASE – *British Academic Spoken English* (Nesi &

Thompson 2006), sviluppato presso le Università di Warwick e Reading. In esso sono contenute le trascrizioni di 160 lezioni frontali (*lecture*) e 40 seminari registrati tra il 1998 e il 2005, per un totale di oltre 1.600.000 *token*. I materiali sono suddivisi in quattro settori disciplinari, quali *Arts and Humanities*, *Life Sciences*, *Physical Sciences* e *Social Sciences*, e sono stati trascritti ed etichettati utilizzando un sistema conforme alle linee guida della TEI – *Text Encoding Initiative* (§1.3.4). Il corpus è accessibile da un'interfaccia *online*, dalla quale si possono effettuare ricerche sui materiali trascritti. Per avere accesso alle videoregistrazioni, si deve farne richiesta al *Centre for Applied Linguistics* dell'Università di Warwick.

Se con i progetti MICASE e BASE l'attenzione è stata posta sulla lingua inglese parlata da soggetti nativi, il progetto *ELFA – English as a Lingua Franca in Academic Settings* (Mauranen 2003) si concentra, invece, sull'uso della lingua inglese sempre in ambito accademico da parte di soggetti non nativi come L2 o come lingua *franca*. Questo corpus contiene un milione di parole, ottenute dalla trascrizione di circa 131 ore di registrazioni effettuate presso quattro diverse università finlandesi (Università di Tampere, Università di Helsinki, Tampere University of Technology e Helsinki University of Technology). In totale, sono stati raccolti gli interventi di circa 650 partecipanti di diversa provenienza, parlanti nativi di 51 altre lingue diverse dall'inglese. Un terzo del materiale riguarda eventi comunicativi con un formato interazionale monologico (lezioni frontali e comunicazioni), mentre la parte restante comprende eventi in un formato dialogico (con l'alternanza tra due o più partecipanti), quali seminari, dibattiti e discussioni di tesi di laurea.

Un tipo simile di varietà linguistica è raccolto nel corpus VOICE – *Vienna-Oxford International Corpus of English* (Breiteneder et al. 2009), contenente un milione di parole trascritte da circa 120 ore di registrazioni. Tuttavia, in questo caso, i 1250 partecipanti coinvolti (tutti parlanti di inglese non nativi) sono stati registrati in situazioni comunicative non accademiche, quali conferenze stampa, interviste, incontri di servizio, vari tipi di dibattiti e discussioni, riunioni e così via. Infine, un ultimo esempio attinente all'inglese come *lingua franca* è il corpus MAW – *Meetings at Work* (Bilbow 2007), nel quale sono raccolti esempi di interazione tra parlanti di inglese non nativi occidentali e parlanti di inglese non nativi di origine cinese (denominati *West expatriates* e *local Chinese* rispettivamente) in un contesto lavorativo ad Hong Kong. Questo corpus è di dimensioni decisamente inferiori rispetto agli altri progetti ELF menzionati: si attesta intorno alle 140.000 parole, tratte da 11 ore di registrazioni nel corso di varie tipologie di riunioni (*departmental management meetings*, *coordination meetings* e *brainstorming meetings*), ognuna con caratteristiche sociolinguistiche proprie (*ibid*., p. 230). Nonostante le piccole dimensioni, questo corpus è stato corredato di un livello di

annotazione particolare e poco diffuso, ossia l'annotazione degli atti linguistici[14] attraverso l'applicazione di etichette individuate appositamente (*ibid*., p. 231), riuscendo in questo modo a esplorare i materiali secondo un paradigma di ricerca prevalentemente (inter)culturale, più che (inter)linguistico.

Nei vari progetti considerati finora, ai quali si potrebbero aggiungere innumerevoli esperienze di ricerca e studio, si riscontrano tutta una serie di strumenti e metodologie, che coincidono o divergono in diversa misura. Inoltre, notiamo che si sono occupati sì di comunicazione parlata, ma soprattutto del tipo di comunicazione con un formato interazionale dialogico (come nella conversazione spontanea, in gran parte delle attività di formazione accademica, o in circostanze costruite *ad hoc* per stimolare la produzione linguistica), monologico (prevalentemente in contesti di natura accademica con eventi comunicativi volti alla formazione e alla trasmissione del sapere), nonché del linguaggio trasmesso attraverso i media (TV, radio, cinema) o il telefono. A ben vedere, pare che siano decisamente di meno gli esempi di ricerche che hanno portato alla realizzazione di corpora orali con materiali tratti specificatamente da convegni o eventi simili. Questo sembra trovare riscontro anche nella panoramica offerta da Bersani Berselli (2004, p. 66) in merito alla situazione dei corpora realizzati con materiali riconducibili alla "conferenza" (sia come tipologia testuale, sia come situazione comunicativa); infatti, sono menzionati uno studio di Webber (1999) e l'archivio di registrazioni realizzate presso la SSLMIT di Forlì. Pur trattandosi di risorse decisamente "minori" per dimensione rispetto agli altri esempi illustrati prima, esse riguardano precisamente il parlato nell'ambito di conferenze e convegni. Tuttavia, è doveroso specificare che solo nel primo caso è lecito parlare di corpus in quanto tale, poiché l'archivio di registrazioni presso la SSLMIT di Forlì è al momento un semplice archivio, cioè una raccolta di materiali che non sono ancora stati strutturati in un corpus elettronico vero e proprio.

Al contributo di Webber sopra citato è possibile aggiungerne un altro da parte della stessa autrice (2004). Questi studi sono di fatto assai utili al fine di definire gli elementi costitutivi della conferenza-convegno (cfr. §4.1), in quanto entrambi sono basati su corpora elettronici e analizzano diversi fenomeni attraverso l'estrazione di liste di frequenza. Nel primo, Webber (1999) ha creato e utilizzato due corpora orali di italiano e inglese ai quali ha affiancato anche due corpora di italiano e inglese scritto (20.000 parole ciascuno) in ambito

[14] Una modalità di studio di alcuni tipi di atti linguistici in un corpus, senza che sia necessario effettuarne previamente una annotazione è suggerita da Kohnen (2000), in riferimento agli «explicit performatives» (*ibid*.,p. 178). Lo stesso sistema potrebbe essere adottato per l'analisi di altri fenomeni di difficile annotazione, di cui una parte limitata potrebbe comunque essere recuperata considerandone i tratti espliciti attraverso i quali sono espressi linguisticamente. Un esempio potrebbe riguardare le metafore su base "un/a specie/sorta/tipo di".

scientifico. I corpora orali riguardano comunicazioni e presentazioni tenute in occasione di convegni, mentre i corpora di lingua scritta raccolgono riassunti di articoli accademici (*abstracts*). L'analisi si concentra sui seguenti elementi: l'uso della deissi pronominale personale, i segnali discorsivi e l'uso dei quantificativi. Purtroppo non abbiamo trovato esplicito riferimento alla metodologia seguita per la costruzione e la consultazione di questi corpora che rappresentano ad ogni modo un materiale estremamente interessante. Nel secondo caso (Webber 2004), il corpus è invece un corpus monolingue (inglese) di interventi tratti da quattro convegni medici internazionali (22.907 parole da *conference monologues* e 10.831 parole da *paper presentations*). Il programma utilizzato per le ricerche semiautomatiche (conteggio parole e estrazione di concordanze) è chiamato *Aston Text Analyser* (sviluppato da Peter Roe). In realtà, di tutti i tratti presi in esame, ovvero «1) passives, 2) informal items, 3) self reference, 4) metadiscourse markers, 5) deixis, modality and instances of humour» (*ibid*., p. 91), non sempre è stato possibile avvalersi di ricerche semiautomatiche. Infatti, «informal items were also counted, but as they consisted mainly of word strings rather than single lexical items, in the event it was found they were detected best by the judicious use of highlighter pens» (*ibid*., p. 92).

A proposito di raccolte di materiali (del tipo conferenza o convegno) registrati e strutturati in archivi multimediali (come è il caso dell'archivio SSLMIT sopra citato), negli ultimi anni sono state messe a punto sempre più risorse di questo genere anche grazie allo sviluppo di Internet. Si pensi non solo alle banche dati di alcune istituzioni internazionali, come la Biblioteca Multimediale del Parlamento europeo e l'archivio multimediale dell'ONU (*UN Webcast Archives*), ma anche ad altre risorse *online*, quali l'archivio *MIT World* del *Massachusetts Institute of Technology*, il portale *Videolectures.net* dell'istituto di ricerca sloveno CT3 e il portale dell'iniziativa TED *Ideas Worth Spreading – Technology, Entertainment, Design* in cui sono disponibili centinaia di videoregistrazioni di conferenze. Nell'insieme, si tratta di risorse eccellenti e con un altissimo potenziale per lo sviluppo di corpora orali (al momento, purtroppo, non forniscono quasi mai la trascrizione dei materiali multimediali).

Uno sviluppo simile ha avuto luogo anche in ambito traduttologico, specialmente a fini pedagogici. Due iniziative considerevoli in tal senso sono il portale *Speech Repository* della DG Interpretazione presso la Commissione europea (con accesso limitato agli istituti membri di questa iniziativa) e la banca dati DAVID – *Digital Audio Video Database* dell'Università di Praga (accessibile solo agli studenti e al personale di detto ateneo). In entrambi i casi, non si tratta di corpora contenenti trascrizioni di dati orali, bensì di banche dati in cui le registrazioni di eventi linguistici disponibili sono state ordinate secondo certi parametri, come illustrato qui di seguito nella Tabella 1.6 e nella Tabella 1.7 rispettivamente:

Tabella 1.6 Parametri di classificazione nella banca dati *EU Speech Repository*.

language	(all official EU languages and candidate-country languages)
level of difficulty	beginner intermediate advanced very advanced
domain	domain of interest divided into EU policy areas (about 30)
intended use	simultaneous consecutive
numerical identifier	(to immediately retrieve a previously identified speech)
title	
type	debate conference press conference hearing pedagogical material
speaker	Surname, Name
details	more information, including speaker's accent, terminology, multimedia (streaming or download) and, occasionally, transcript

Tabella 1.7 Parametri di classificazione nella banca dati *DAVID*.

languages	
source	DVD consecutive simultaneous mock conference
level of difficulty	easy difficult
subjects	
title	
topic	
date and speaker	
file format	
file size	
duration	
short description	(briefing)
keywords	

Uno strumento a metà strada, in un certo senso, tra quest'ultimo tipo di banche dati e un corpus vero e proprio è la banca dati *Marius* (de Manuel 2003b), sviluppata presso l'Università di Granada in Spagna. *Marius* contiene registrazioni di interventi individuali (TP) di varia natura, i quali sono stati registrati sia dal canale satellitare *Europe by Satellite* (dibattiti in seno al Parlamento europeo e fora tematici sul multilinguismo), sia direttamente sul

campo (convegni specialistici tenutisi in Spagna, Forum Sociale Mondiale e Forum Sociale Europeo). Una buona parte dei materiali è accompagnata dalle trascrizioni, redatte secondo convenzioni che hanno tenuto conto dell'uso pedagogico dei materiali (per esempio, vi sono segnalazioni dei nomi propri, normalizzazioni di eventuali disfluenze, ecc.). Nonostante questa scelta vada a inficiare eventuali analisi sulle caratteristiche linguistiche specifiche di questi TP, essi sono stati catalogati accuratamente e sono pertanto arricchiti di numerose informazioni metatestuali, le quali consentono di estrapolare caratteristiche rilevanti sul tipo di intervento, oratore e altro ancora. La classificazione di tutti i materiali è stata basata su 31 diversi attributi, suddivisi in cinque grandi categorie:[15]

> • *location fields*: code, title, reference of the tape, start and end cues of the speech within a larger recording as well as corpus and sub-corpus affiliation,
> • *descriptive objective fields*: date; name and position of the speaker; language, duration, number of words and average delivery rate of the speech; additional information such as time limitations imposed on the speaker, as in the case of European Parliament speeches;
> • *approximate or subjective data fields*: level of specialization, speaker's accent, 'time autonomy' and 'context autonomy' of the speech;
> • *pedagogical orientation fields*: indications of the training stage and interpreting modality for which each piece of material is recommended (e.g. initial consecutive, intermediate simultaneous, simultaneous with text, etc.);
> • *objects* (transcripts) and hyperlinks to clips, related documents or relevant websites.
> (Sandrelli & de Manuel 2007, p. 281)

Oltre a fornire una descrizione generale dei vari materiali raccolti nella banca dati, tutte queste informazioni testuali e metatestuali attribuite a ciascuna registrazione e trascrizione risultano di estrema utilità a fini didattici nella formazione degli interpreti. A questo proposito, i vantaggi che vi si possono trarre sono molteplici: è possibile proporre materiali autentici fin dai primi momenti della formazione in aula; si possono selezionare i materiali secondo

[15] Le stesse sono state formulate prima ancora in lingua spagnola:
> […] *datos de localización* (para ubicar cada discurso en la cinta VHS, el CD o el DVD en el que está grabado); *datos descriptivos objetivos* (nombre y condición del orador, idioma, duración, número de palabras, velocidad de elocución, fecha, entre otros); *datos valorativos o aproximativos* (nivel de especialización, acento, autonomía temporal y contextual, entre otros); datos de orientación pedagógica (derivados de los anteriores); *objetos OLE* (textos de las transcripciones de los discursos con anotaciones didácticas y clips de vídeo en formato AVI); e *hipervínculos* con enlaces a Internet o a documentos propios del ordenador sede.
> (de Manuel 2003b, p. 37)

diversi livelli di difficoltà, creando quindi un percorso graduale per lo sviluppo e l'acquisizione di competenze individuali (Kalina 2000); la valutazione è resa più agevole dalla disponibilità delle trascrizioni; in generale, la banca dati può essere utilizzata dagli interessati anche come risorsa per l'auto-apprendimento[16] e per la formazione continua.

A questo punto, risulta superfluo puntualizzare che simili strumenti, siano essi corpora o banche dati, richiedono un ingente investimento di risorse umane, intellettuali (interdisciplinari) e finanziarie. Eppure, una volta messi a disposizione della comunità scientifica, diventano risorse straordinarie da usare per innumerevoli studi, attività didattiche e come termine di paragone. Nonostante la validità di questa affermazione, la vastissima gamma di varietà e comunità linguistiche del comunicare umano solleva non poche questioni di rappresentatività. Per questo, talvolta è insufficiente, se non del tutto impossibile, sfruttare risorse già esistenti, poiché queste potrebbero non essere pienamente pertinenti con quanto si intende studiare, così come potrebbero essere state realizzate secondo criteri che si discostano troppo dal nostro interesse. Per tali motivi, alle tipologie di corpora illustrate all'inizio di questo capitolo si potrebbe aggiungere anche il corpus "D.I.Y.", ovvero *do-it-yourself* (McEnery et al. 2006, pp. 71 e ss.; Maia 1997), a cui appartengono anche i due corpora che saranno presentati nel corso del presente lavoro. Oltre ad essere due *D.I.Y. corpora*, EPIC e DIRSI-C sono corpora di lingua parlata (e tradotta simultaneamente), multilingue, paralleli e comparabili, annotati; per quanto riguarda DIRSI-C, esso è anche allineato (sia con un allineamento testo-suono, sia TP-TA sulla base del contenuto). Si tratta pertanto di due risorse linguistiche che sono frutto di una proposta di applicazione della Linguistica dei Corpora (*Corpus Linguistics*) agli Studi sull'interpretazione (*Interpreting Studies*), con l'auspicio che possa contribuire alla crescita dei *Corpus-based Interpreting Studies* (CIS).

1.3 Sfide metodologiche nei CIS

Dalla definizione di corpus precedentemente presentata (§1.1), si intuisce che vi sono diverse tappe nella realizzazione di un D.I.Y. (*do-it-yourself*) corpus. Ogni tappa comprende un certo numero di attività, le quali si differenziano a seconda del tipo specifico di corpus che si intende costruire (Tabella 1.1, Tabella 1.2; Sinclair 1991, pp. 13-23; McEnery et al. 2006, pp. 71-79). L'intero procedimento è altresì accompagnato da ostacoli e sfide metodologiche,

[16] Sulle applicazioni informatiche pensate per l'auto-apprendimento nella formazione degli interpreti si vedano, tra gli altri, Sandrelli (2002, 2003a, 2003b) e Carabelli (2003).

anch'esse variabili a seconda del tipo di corpus in questione, ma che in sostanza sembrerebbero pesare particolarmente nell'ambito dei CIS. Infatti, le difficoltà generalmente riconosciute in merito alla creazione di corpora di lingua parlata (rispetto ai corpora di lingua scritta) sono per certi versi raddoppiate nei CIS, poiché la comunicazione non è solo parlata, ma anche interpretata. Tale duplicità si riflette direttamente sulla conformazione della comunità linguistica di riferimento (tra l'altro nettamente ristretta rispetto ad altre forme del comunicare umano), nonché sul numero di variabili che accompagnano la produzione della varietà linguistica (o meglio, varietà linguistiche, dato che si ha a che fare con lingua "originale" e lingua "interpretata", oltre che con linguaggi settoriali) oggetto di studio.

Partendo ancora una volta dalla considerazione che i corpora di interpretazione sono anzitutto corpora di lingua parlata, ai fini della loro realizzazione sarà necessario percorrere per lo meno tutte le tappe previste nella compilazione di uno *spoken corpus*. Esse sono state riassunte nel seguente modo (Thompson 2005), ma andranno integrate per sfruttare al meglio le potenzialità di ricerca tipiche dei CIS:

Tabella 1.8 Tappe fondamentali nella creazione di un corpus orale.

1	Data collection
2	Transcription
3	Markup and annotation
4	Access

Prima ancora di procedere con la fase iniziale di raccolta dei dati, è in realtà necessario circoscrivere l'oggetto di studio per poter determinare la *dimensione* e la *struttura* del corpus (§1.3.1). In altre parole, la raccolta effettiva dovrebbe essere mirata a del materiale che risulta essere pertinente secondo alcuni criteri di rappresentatività già individuati; per esempio, la modalità di interpretazione e il contesto, così come il tipo di interazione e di eventi linguistici. In questa fase, è di importanza fondamentale assicurarsi che non vi siano restrizioni all'uso dei dati raccolti per questioni di riservatezza e copyright (§1.3.2.2). Trattandosi di dati orali, la raccolta va effettuata attraverso la *registrazione* audio o video, attualmente resa più agevole rispetto al passato grazie ai progressi straordinari della tecnologia in materia di registrazione e successiva visualizzazione dei dati (§1.3.2.3). La fase successiva comporta la *trascrizione* dei dati registrati, da effettuare sulla base di convenzioni coerenti e assicurandone la leggibilità sia da parte del trascrittore/ricercatore, sia da parte del computer/programma informatico (Edwards & Lampert 1993, §1.3.3). Dopo aver trascritto i dati, si passa alla loro *codifica* e *annotazione* (§1.3.4), operazioni che consentono di

aggiungere informazioni di vario tipo all'interno del corpus, potendo così effettuare ricerche semiautomatiche mirate e altrimenti non realizzabili manualmente se si lavora su grandi quantità di dati. Un esempio particolare di annotazione si può ottenere con il *Part-of-Speech tagging* (*POS-tagging*), cioè l'assegnazione a ciascun *token* di un'etichetta con informazioni di natura grammaticale e morfologica (§1.3.4.1). Va precisato che esistono più tipi di annotazione che possono riguardare anche altri livelli, quali ad esempio il livello paralinguistico, metalinguistico e così via. Sempre in questa fase, un'ulteriore operazione che darebbe un notevole valore aggiunto al corpus riguarda l'*allineamento testo-suono/video*, cioè dei testi trascritti con le rispettive registrazioni audio/video; nel caso specifico dei corpora orali paralleli (realizzabili, per l'appunto, con materiali risultanti da situazioni comunicative mediate da interpreti), si potrebbe effettuare anche un *allineamento TP-TA*, sulla base del contenuto e/o del *décalage* a seconda della modalità (§1.3.5). Infine, l'ultima fase consiste nel rendere possibile l'*accesso al corpus*, mettendolo quindi a disposizione della comunità scientifica interessata e, possibilmente, corredandolo di strumenti di ricerca semiautomatica, per esempio attraverso interfaccia *online* e programmi informatici dedicati (§1.3.6).

Da questa breve disamina, le tappe fondamentali nella creazione di un corpus di interpretazione possono essere così riassunte (Tabella 1.9):

Tabella 1.9 Tappe fondamentali nella creazione di un corpus di interpretazione.

1	*Corpus design*
2	*Data collection*
3	*Transcription*
4	*Markup and annotation*
5	*Alignment*
6	*Access*

Per ciascuna delle varie tappe schematizzate nella Tabella 1.9, fin dagli albori dei CIS è stata messa in risalto una parte sostanziale delle sfide metodologiche che ne hanno probabilmente rallentato lo sviluppo rispetto ai più avanzati CTS. Nonostante la straordinaria evoluzione dei CIS a cui si è assistito negli ultimi anni (§2), Setton (*s.d.*) li identifica, non a torto, come una sorta «cottage industry» (*ibid.*), ovvero un settore di nicchia in cui pochi hanno tentato finora di avventurarsi. Tuttavia, come già sottolineato più volte, questo non è certo dovuto a una mancanza di interesse da parte dei ricercatori.

Uno dei primi studiosi ad evidenziare i principali ostacoli al *corpus-based approach* è stata Armstrong (1997), la quale solleva anzitutto la questione della quantità e del tipo di dati disponibili, poiché «A prerequisite for corpus-based

studies is the availability of adequate electronic data» (*ibid*, p. 150). Per capire in che misura i dati siano "adeguati", la stessa Armstrong suggerisce di basarsi su un certo numero di criteri che sono, al tempo stesso, potenziali fonti di sfide: la qualità tecnica, il tipo di lingua, la rappresentatività, le possibilità di annotazione, nonché i diritti d'uso e di distribuzione dei dati stessi. Shlesinger (1998) insiste particolarmente sulle difficoltà legate alla trascrizione dei dati orali e all'effettiva possibilità di rappresentarne i tratti paralinguistici: «The difficulty lies not only in the act of transcription, per se, but in the fact that certain elements of spoken communication are both so subtle and so subjective as to defy description […]» (*ibid*., p. 487). Su quest'ultimo punto è dello stesso avviso anche Cencini (2002), il quale ha approfondito il tema degli ostacoli derivanti dall'ottenimento, dalla produzione delle trascrizioni e dalla loro codifica e annotazione. Questi possono essere così sintetizzati: le trascrizioni sono difficili da ottenere; trascrivere dati orali è un'attività che richiede molto tempo; le trascrizioni sono rappresentazioni parziali dei dati interessati; non esiste uno standard generale per le convenzioni di trascrizione e, conseguentemente, i dati non possono essere condivisi e scambiati tra diversi studi; gli strumenti di analisi sono limitati.

È evidente che le competenze richieste a chi si prefigge l'obiettivo di percorrere tutte le tappe sopra elencate sono marcatamente multidisciplinari: dalla Linguistica alla Traduttologia, dagli studi sulla Comunicazione Parlata all'Informatica. Soprattutto per quest'ultimo settore disciplinare, il più delle volte il ricercatore-interprete o *practisearcher* avrà necessità di avvalersi di un aiuto esterno; in termini pratici, questo significa che occorrerebbe avviare un rapporto di collaborazione con altri soggetti (ad esempio, esperti linguisti computazionali, tecnici informatici, laboratori), mettendo in campo il meglio delle proprie capacità relazionali. La questione è più delicata di quel che si pensi, poiché pur riuscendo a far dialogare tra loro diversi settori disciplinari, non è detto che si riesca a garantire fin da subito un buon livello di reciproco intendimento:

> You do not need to be a jack-of-all trades to become a corpus linguist but you may well find yourself, after years in the field, having had to learn a bit of everything. Alongside the theoretical challenges, corpus linguists fight many practical battles with text editors, concordance tools, mark-up format and linguistic annotation. For anyone aiming to compile their own corpus, the endeavours of cleaning up text, tokenising, indexing, and annotating can become an insurmountable task, especially if they are to be done manually. If manual labour is not an option, particularly when working with very large corpora, then another option is often to wait for your organisation's over-worked technical staff to allocate time for investigating the problem. This usually involves some more or less embarrassing moments of mis-

> communication, where the power is with the one who knows which tasks are impossible to implement, while other tasks are achievable almost instantaneously. For the technical virgin, the difference between an impossible or possible computing task is often hidden in darkness. Why should it be so easy to count frequencies when it is so difficult to mark sentences?
> (Danielsson 2004, p. 225)

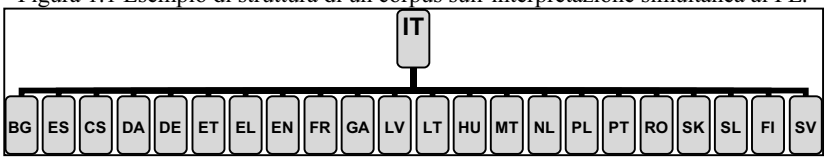
Nelle sezioni successive ripercorreremo ogni tappa fondamentale della costruzione di un corpus di interpretazione, fornendo di volta in volta possibili soluzioni alle *theoretical challenges* e alle *practical battles* che accompagnano questa esperienza di ricerca (§2). Nei capitoli seguenti (§3 e §4) saranno invece illustrate le soluzioni adottate specificatamente nella creazione del corpus EPIC e del corpus DIRSI.

1.3.1 Corpus Design

1.3.1.1 Struttura del corpus

La struttura di un corpus di interpretazione è direttamente determinata dal numero di lingue coinvolte, dalla modalità di interpretazione e dal tipo particolare di corpus che si intende realizzare (parallelo, paragonabile, ecc.). Nel caso specifico dell'interpretazione simultanea, si avranno uno o più sottocorpora di TP abbinati a uno o più sottocorpora di TA. Per esempio, se si considerano le attuali sedute plenarie del Parlamento europeo, si potrebbe ipotizzare un corpus di interpretazione dell'italiano. La sua struttura comprenderebbe un sottocorpus contenente discorsi originali in italiano e ventidue sottocorpora contenenti i TA, cioè le interpretazioni nelle altre lingue ufficiali dell'Unione europea, come rappresentato graficamente nella Figura 1.1, dove i vari sottocorpora sono indicati dalla sigla delle lingue in uso:

Figura 1.1 Esempio di struttura di un corpus sull'interpretazione simultanea al PE.



Lo stesso schema andrebbe ovviamente ripetuto per ciascuna lingua, ottenendo così altrettanti sottocorpora di TP e TA. Semplificando la struttura a due lingue di lavoro (italiano e inglese), i quattro sottocorpora risultanti (TP-IT e TP-EN con TA-EN e TA-IT) possono essere strutturati in modo parallelo o comparabile. Per esempio, si potrebbero selezionare i discorsi originali in italiano e i rispettivi TA in inglese, ma anche esaminare le caratteristiche dei discorsi originali italiani rispetto alle interpretazioni in italiano dei TP inglesi: stessa lingua (l'italiano), ma condizioni di produzione differenti. Oltre a questo, se alle trascrizioni fossero abbinate le registrazioni audio o video, il corpus acquisirebbe un carattere multimediale, con possibili conseguenze sull'architettura interna (informatica). Lo stesso si potrebbe dire se si volessero allegare eventuali documenti o supporti audiovisivi utilizzati dagli oratori, operazione che avvicinerebbe il corpus alla tipologia multimodale. Infine, anche l'eventuale ottenimento di più TA da uno stesso TP andrebbe ad incidere sull'organizzazione della struttura del corpus. Si potrebbero avere più TA prodotti nella stessa modalità (per esempio con più cabine della stessa lingua in servizio contemporaneamente, o in più sedute sperimentali), così come più TA prodotti secondo modalità diverse (simultanea con o senza *relais*, consecutiva, traduzione scritta, ecc.), da cui si otterrebbe un corpus intermodale. In questo caso, le diverse condizioni di produzione e la natura specifica dei singoli TA potrebbero portare a soluzioni alternative su come gestire l'architettura generale del corpus.

1.3.1.2 Rappresentatività

La questione della rappresentatività potrebbe essere riassunta nelle seguenti domande: di che cosa ci vogliamo occupare? Quale comunità linguistica e quale varietà linguistica ci interessa studiare? Che tipo di situazioni comunicative intendiamo circoscrivere e analizzare?

Per rispondere a tali quesiti è opinione condivisa che sia necessario svolgere anzitutto un'approfondita riflessione teorica. Questo è valido tanto per la *Corpus Linguistics* (con le sue applicazioni alla traduttologia, quali i CTS e i CIS), quanto per qualsiasi altra disciplina che si prefigge di descrivere una realtà oggetto di studio.

Tra i tanti approcci alla descrizione della lingua e della comunicazione, possiamo innanzitutto ispirarci agli strumenti da sempre usati nella Sociolinguistica per affrontare il tema della rappresentatività. A questo proposito, un avvertimento essenziale è di non trarre conclusioni generali e assolute dall'analisi di un gruppo parziale della popolazione oggetto di studio. Ovviamente, l'ingrediente chiave per ottenere una solida rappresentatività dei dati sarebbe evitare il più possibile ogni tipo di *bias*. Tuttavia, si riscontreranno

sempre dei limiti di rappresentatività in un modo o nell'altro (Milroy & Gordon 2003). La riflessione teorica dovrebbe, quindi, prendere le prime mosse da una concettualizzazione del campione da analizzare, per la quale Sankoff (1980a, p. 26) individua le seguenti tre fasi:

1. definire l'universo di campionamento;
2. valutare le dimensioni della variazione all'interno della comunità;
3. determinare la dimensione del campione.

In genere, la definizione dell'universo di campionamento (*sampling universe*) non rappresenta un problema, ma va determinata tenendo conto di una considerazione logistica importante, nel senso che bisogna sapere come avere accesso alla popolazione target. Nella seconda fase, si dovrebbero stabilire le dimensioni della porzione di universo campione in cui sono presenti le variazioni e i tratti rilevanti alla ricerca in questione. Negli studi di natura linguistica, i campioni sembrano essere sempre piuttosto ridotti rispetto a quanto indicato dalla teoria. Teoricamente, un campione pari a 300 sarebbe idoneo per una popolazione piccola (inferiore a 1.000), mentre una popolazione grande (oltre 150.000) potrebbe essere rappresentata da un campione pari a 1.500 (Neuman 1997, p. 222). Per gli studi linguistici la situazione è solitamente lontana da questi parametri, ma si ritiene anche che non sia assolutamente necessario avere un campione tanto cospicuo, in quanto l'uso della lingua tenderebbe a essere più omogeneo di altri fenomeni studiati in altre discipline (Labov 1966, pp. 180-181).[17]

Entrando nello specifico degli studi che prevedono la creazione di corpora linguistici, Biber (1993) descrive egregiamente il quadro generale delle operazioni che sottostanno all'ottenimento di un adeguato grado di rappresentatività. Questo stesso contributo è ripreso da Halverson (1998), la quale lo applica ai CTS e conferma che «all discussion of corpus text selection and classification, the types of analysis adopted, and the significance of the findings must be grounded in an explicit description of what the enquiry takes to be its object» (*ibid*., p. 595/2). Da queste due trattazioni, le tre fasi sopraelencate assumono contorni più precisi e sono corredate da strumenti specifici.

In linea con quanto affermato poco sopra, una prima osservazione è che la dimensione del campione non è il parametro essenziale su cui tarare il grado di rappresentatività (questo è a dir poco rassicurante, se si pensa alla situazione degli studi complessivamente condotti nel campo dell'interpretazione). Si tratta, dunque, di definire al meglio la popolazione oggetto di studio o *target*

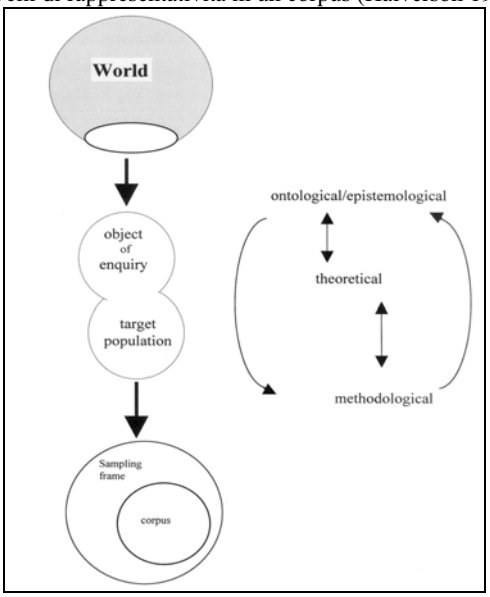
[17] A questo si aggiungono anche i problemi legati alla gestione dei dati raccolti (§1.3.2.3.2), problemi che spesso spingono i ricercatori a ridimensionare il campione effettivamente studiato o, per lo meno, a selezionare una parte di quanto è stato raccolto (Kalina 1994).

population, in particolare circoscrivendone i confini (criteri di inclusione ed esclusione) e stabilendone l'organizzazione gerarchica interna. Queste due operazioni sono realizzabili a fronte di una previa riflessione di tipo teorico sull'oggetto di studio e sulla finalità del corpus, dopodiché esse possono essere effettuate seguendo diversi approcci e parametri. Il passaggio dal piano teorico a quello applicativo è accompagnato dalla trasformazione del costrutto definito come "popolazione target" in una lista di campionamento vera e propria, chiamata "*sampling frame*", ovvero «an itemized listing of population members from which representative samples can be chosen» (Biber 1993, p. 224). In sintesi:

> For the purpose of corpus construction, the conception of the object which a discipline more or less agrees on provides the motivation for defining a target population. The specification of a target population then provides the basis for selecting a sampling frame. The sample taken from that frame constitutes the corpus, which is thus a representation of a larger set of phenomena.
> (Halverson 1998, p. 498/5)

Il tutto avviene secondo un processo lineare, ma al tempo stesso dinamico, in cui «the insight gained at one level of enquiry will have an effect on the other levels as well» (*ibid.*, p. 498/5), anche in base alla metodologia adottata. L'intero processo è illustrato nella Figura 1.2 sotto riportata:

Figura 1.2 Livelli di rappresentatività in un corpus (Halverson 1998, p. 498/5).



Volendo applicare il processo definitorio sopra illustrato al mondo della Traduzione, vi sono due fattori particolarmente rilevanti di cui si dovrebbe tenere conto (*ibid.*). Il primo è la distinzione tra *natural* e *professional translators/interpreters*, poiché la mediazione interlinguistica è un'attività che da sempre accompagna l'esistenza stessa dell'umanità ed è esercitata sia in modo "spontaneo" e "naturale" da chiunque conosca almeno due lingue, sia da professionisti che ne hanno fatto il loro mestiere, anche a seguito di uno specifico percorso di formazione (laurea, master, ecc.). Il secondo fattore è la direzionalità: si considera se l'interprete traduce dalla lingua straniera verso la propria lingua madre (cioè, secondo la classificazione proposta dall'Associazione Internazionale Interpreti di Conferenza, dalla lingua B o C alla lingua A) o, viceversa, dalla propria lingua materna verso una lingua straniera (dalla lingua A alla lingua B).

Dopo aver impostato i confini della popolazione target con questi due fattori, le altre variabili da considerare, al fine di strutturarne la gerarchizzazione interna, potrebbero essere raggruppate in due grandi categorie, ovvero *subject variables* e *discourse variables* (Setton 2002). Tra le variabili legate ai soggetti, Setton (*ibid.*) menziona il grado di formazione e il livello di esperienza degli interpreti, così come la loro preparazione per un ingaggio specifico (quest'ultima più ostica da determinare ed esprimere in unità discrete); dall'altra parte, le variabili legate ai testi e agli eventi linguistici comprendono la velocità di eloquio, il grado di tecnicità e spontaneità, il genere e il registro. A queste si aggiungono le specificità dovute a particolari abbinamenti linguistici e alle condizioni di lavoro (per esempio, la disponibilità del testo scritto durante la lettura di un discorso in simultanea), nonché le informazioni sul contesto.

Complessivamente, notiamo che i fattori e le variabili menzionate finora sono applicabili ora al TP (e al soggetto che lo produce), ora al TA (e ai traduttori/interpreti), e il più delle volte a entrambi. A ben vedere, sembrerebbe opportuno operare uno sdoppiamento del processo rappresentato nella Figura 1.2, in quanto stiamo ragionando su due mondi interdipendenti. Ai due mondi afferiscono una popolazione target di TP e una popolazione target di TA, entrambe personificate da soggetti con un proprio ruolo comunicativo: scrittori e oratori da una parte (popolazione target della LP), traduttori e interpreti dall'altra (popolazione target della LA).

È interessante notare come questa visione di duplice sdoppiamento sia in linea con quanto affermato da Meyer & Nelson (2006, p. 107): «Sampling frames are used extensively in designing written corpora, but cannot be applied to most kinds of spoken data. Instead, demographic sampling is used». Nel caso specifico dei corpora di interpretazione, la scelta su quale popolazione target considerare potrebbe spaziare lungo la serie di modalità e contesti di

interpretazione.[18] Tuttavia, uno dei maggiori problemi nel riuscire ad applicare qualsiasi tassonomia basata su un repertorio di categorie prestabilite è dato dalla rigidità dei sistemi di classificazione. Potrebbero verificarsi casi ambigui o ibridi, infatti, laddove un membro di una determinata categoria presenti caratteristiche compatibili anche con altre categorie. Per rendere il tutto più flessibile, Halverson (1998) propone di basarsi su un sistema di categorie prototipiche, in cui «the boundary is not fixed, and the members are not equal» (*ibid*., p. 507/14). Il vantaggio risiederebbe nel fatto che «a prototype category structure provides us with […] a means of addressing the relativity of definitions, and of studying various kinds of translation within an integrated framework. It also provides us with a means of coping with the quite obvious asymmetries in the category members» (*ibid*., p. 511/18). In questo senso, l'etichetta "interpretazione simultanea" è a tutti gli effetti applicabile a condizioni di lavoro anche molto diverse tra loro (con o senza cabina insonorizzata, con o senza supporti scritti, con o senza condivisione dello spazio in cui avvengono gli scambi comunicativi tra i partecipanti primari e secondari, e così via). Il profilo degli stessi interpreti è soggetto a variazioni; oltre alle tipologie menzionate in questa sezione, sarebbe lecito considerare una distinzione anche tra liberi professionisti e interpreti funzionari, o ancora tra interpreti con maggiore o minore esperienza (una categoria, certamente, di non facile definizione in quanto l'esperienza dipenderebbe non solo dal numero di ingaggi, ma anche dalla loro varietà e distribuzione nel tempo, dal grado di complessità, ecc.). Allo stesso modo, per quanto riguarda la situazione comunicativa, sono tante le forme che possono assumere le *stage activities* mediate da interpreti, tra cui lo stesso concetto di conferenza si presta a più letture da diverse prospettive (l'evento generale o la presentazione di una relazione).

Sul tema della rappresentatività in riferimento specifico al nostro oggetto di studio, la conferenza-convegno, Bersani Berselli (2004, pp. 63-68) mette in luce varie questioni metodologiche di particolare interesse, quali «la definizione dei limiti entro cui si debba considerare un testo come esemplare autentico di "conferenza" e quindi l'articolazione del genere in sottogeneri, ed eventualmente in sotto-sotto-generi» (*ibid*., p. 64); la considerazione del relativo oggetto o campo disciplinare a cui appartiene la conferenza; infine, la popolazione di riferimento, nonché i tratti linguistici che si intende radunare

[18] Consideriamo qui la distinzione proposta da Hurtado (1994/1995, 1996) e Jiménez Ivars (1999, 2002) tra modalità di interpretazione in simultanea e in differita, a seconda della contemporaneità o meno della produzione di TP e TA. Tra le modalità in simultanea rientrano la simultanea con o senza cabina insonorizzata, il *chuchotage* e la traduzione a vista; le modalità in differita comprendono la consecutiva standard con presa di note e la consecutiva breve (prevalentemente senza presa di note). Le varie modalità considerate all'interno di contesti comunicativi specifici danno luogo a diversi *tipi* di interpretazione, quali l'interpretazione di conferenza, l'interpretazione di trattativa, l'interpretazione giuridica e così via (Pöchhacker 2004, pp. 13-16).

assieme. Tali tratti si riferiscono alle caratteristiche peculiari dei testi che si vogliono analizzare, tra cui quelli citati da Bersani Berselli riguardano il grado di pianificazione e il grado di interattività. Anche in questo approccio, l'individuazione di generi e sottogeneri conferma la seconda operazione presentata da Biber (1993) e discussa da Halverson (1998), ovvero l'identificazione di una strutturazione gerarchica interna della popolazione di riferimento «through the specification of so-called strata, or sub-categories» (*ibid.*, p. 503/10).

Da ultimo, vale la pena rammentare che le questioni di natura più pratica nella raccolta dei dati per i CIS concorrono sostanzialmente a determinare la dimensione e la "fisionomia" del campione studiato. Nelle sezioni successive approfondiremo i maggiori punti di potenziale criticità che si possono riscontrare durante la raccolta dei dati, cioè la seconda tappa prevista nella compilazione di un corpus di interpretazione.

1.3.2 Raccolta dei dati

Tra le considerazioni sulla rappresentatività di un corpus esposte nella sezione precedente, è stato fin da subito puntualizzato che la definizione dell'universo di campionamento non può prescindere dalla conoscenza di come avere accesso alla popolazione target. Se, come abbiamo visto, nel caso dei corpora di interpretazione si ha uno sdoppiamento della popolazione target, nel nostro caso specifico dovremo avere accesso sia ai partecipanti primari, sia agli interpreti, così come si dovranno impiegare strategie di raccolta dei loro TP e TA rispettivamente. Pertanto, le sfide metodologiche generali su cui concentreremo la nostra attenzione sono le seguenti: l'accessibilità alla comunità linguistica oggetto di studio (oratori e interpreti) e ai loro TP e TA; la registrazione dei dati orali (a cui si accompagna talvolta la raccolta e il reperimento dei materiali di supporto utilizzati nel corso della trasmissione dei TP); la gestione dei dati raccolti. Ogni sfida generale presenta molteplici sfaccettature che verranno approfondite nelle sezioni a seguire.

1.3.2.1 Accessibilità

Una delle prime questioni che si devono affrontare nella creazione di un corpus riguarda il grado di accessibilità ai dati che sono l'oggetto di interesse per lo studio. La stessa questione è certamente valida per molti altri campi di ricerca, ma negli Studi sull'interpretazione (quale che sia la modalità considerata) merita particolare attenzione. Da sempre si lamenta che gli Studi sull'interpretazione si

sono basati su campioni di dati molto limitati, per non dire esigui o aneddotici (Shlesinger 1998); questo è comprensibile se si pensa alle tante variabili in gioco che rendono ciascun evento mediato da interpreti un evento dalle caratteristiche uniche e irripetibili. Oltretutto, le difficoltà di accesso ai dati rappresentano un ostacolo anche alla descrizione delle situazioni comunicative stesse in cui operano gli interpeti, come è stato ribadito in un articolo "denuncia/appello" (Pöchhacker 2009) con particolare riferimento allo studio dell'interpretazione in ambito sociosanitario e giuridico.

Essendo comunque auspicabile che un corpus contenga una grande quantità di dati (cioè, semplificando, un numero sufficientemente rappresentativo di TP e TA registrati e trascritti), Armstrong (1997) suggerisce di tenere conto che «Large collections of homogeneous data […] are typically found either in large international political organizations or represent technical documentation produced by major multinational companies» (Armstrong 1997, p. 151). Tralasciando il contesto delle multinazionali e la loro produzione di documentazione tecnica (un contesto dove spesso agli interpreti ingaggiati è richiesto anche di firmare un accordo di riservatezza vincolante), le sedi delle organizzazioni internazionali rappresentano effettivamente una delle fonti più preziose di dati per i nostri scopi, sia rispetto ai documenti scritti, sia per quel che riguarda i servizi di interpretazione. Basti ricordare le assemblee dell'ONU e della NATO, le istituzioni dell'Unione europea e gli organismi pubblici presenti in realtà bilingui come il Canada, oppure in realtà caratterizzate dal multilinguismo come il Sudafrica e la Malesia. Tra tutti, le sedi parlamentari presentano forse un grado di accessibilità maggiore (Ibrahim 2009), in considerazione del principio di trasparenza su cui si fondano generalmente le attività che vi si svolgono. Ad ogni modo, non va sottovalutato che l'accessibilità è parimenti determinata dalle circostanze fisiche e dalle persone direttamente coinvolte nella ricerca.

In generale, la continua evoluzione della tecnologia, con la disponibilità dei sistemi di trasmissione e registrazione digitali, gli strumenti di registrazione (video e audio) nonché di supporto e di trasmissione dei dati (internet, banda larga, hard disk portatili), apre costantemente nuove possibilità di accesso a una gamma sempre più vasta di situazioni comunicative (de Manuel 2003b, pp. 27-35). Ad esempio, riprendendo i casi a cui si è accennato poco sopra, gran parte delle principali istituzioni internazionali mette a disposizione in Rete le registrazioni di molti eventi comunicativi, scaricabili anche in un PC. La Biblioteca Multimediale del Parlamento europeo ne è un esempio notevole, in quanto mette a disposizione non solo il video dei TP ma anche le registrazioni dei TA in tutte le lingue ufficiali dell'Unione per ogni seduta plenaria.[19]

[19] I primi materiali contenuti in questo archivio multimediale risalgono alle sessioni tenute nel mese di aprile 2006.

Precursore di questo straordinario sistema di divulgazione delle attività in seno al PE è stata la televisione satellitare. Molto prima che fosse messa a punto la Biblioteca Multimediale del PE, il canale *Europe by Satellite* (EbS) ha puntualmente trasmesso, e trasmette tuttora in diretta, parte dei dibattiti tenuti durante le plenarie del PE (così come diverse conferenze stampa e altri eventi comunicativi) con la possibilità di sintonizzare il canale di ascolto sull'originale (*floor*) o su ognuna delle cabine degli interpreti in servizio. Questi canali di accesso ai dati "dall'esterno" sono stati già utilizzati in importanti progetti di ricerca, tra cui la banca dati *Marius* (de Manuel 2003b) e lo stesso corpus EPIC (§3), e continueranno a rappresentare senza dubbio una via particolarmente vantaggiosa per servirsi di una fonte inesauribile di dati preziosi, quale è il Parlamento europeo (Bendazzoli 2010).

Il contesto delle istituzioni UE e la comunicazione mediata da interpreti simultanei che lo caratterizza sono stati studiati accedendo ai dati anche "dall'interno". In uno studio su un campione di 120 TP in quattro lingue (inglese, finlandese, tedesco e svedese) e dei relativi TA in tutte le lingue coinvolte, Anna-Riitta Vuorikoski (2004) ha avuto la possibilità di avvalersi dell'aiuto dello staff tecnico del Parlamento europeo per reperire i dati, probabilmente coadiuvata dal fatto di prestare lei stessa servizio come interprete per le istituzioni comunitarie:

> The initial impulse for this study came while I was employed as an interpreter for the EP. This gave me ample opportunity to make observations about the speech context. As certain elements of the EP genre became increasingly dominant for the research plan it seemed advisable to interview a few Finnish MEPs for their views of the way in which the meetings function.
> (Vuorikoski 2004, pp. 92-93)

È evidente come il fatto di non essere un completo *outsider* abbia contribuito non solo a ispirare e orientare lo studio di Vuorikoski, ma anche ad avere un accesso più agevole ai dati (cioè alle persone coinvolte nella situazione comunicativa studiata, nonché al personale tecnico dei servizi audiovisivi del PE da cui ha ricevuto assistenza nella fase di selezione delle registrazioni).

Il panorama cambia decisamente se ci spostiamo nel mercato privato italiano (e probabilmente di molti altri paesi), dove gli ostacoli che si frappongono alla raccolta dei dati, laddove questi sono registrazioni di oratori e di interpreti, sono decisamente più numerosi (Kalina 1994, Zorzi 2004). Innanzitutto, vi è una differenza a seconda della natura pubblica o privata dell'evento in esame. Possibili restrizioni potrebbero essere dovute alla riservatezza delle informazioni scambiate dagli interlocutori, pertanto dovrebbe risultare più facile accedere ad eventi pubblici, soprattutto a quelli pensati

proprio per avere un'ampia diffusione di ciò che è presentato. In entrambi i casi è comunque buona prassi fare uso di un consenso informato (§1.3.2.2), così come si rende necessario vagliare tutte le caratteristiche della situazione comunicativa che si vuole osservare. La riflessione teorica sul contesto in cui ha luogo la comunicazione mediata dagli interpreti, sulle dinamiche interne degli eventi comunicativi, i diversi partecipanti coinvolti e i tipi di eventi linguistici scambiati tra loro dovrebbe orientare, almeno in parte, i primi passi verso l'accesso ai dati. Nel caso dei convegni internazionali, i potenziali interlocutori sarebbero senz'altro gli interpreti, i partecipanti all'evento, gli organizzatori e, eventualmente, gli iniziatori (Russo 1999). Oltre a fare affidamento ai contatti diretti e personali a disposizione di ogni ricercatore, ci si potrebbe dunque rivolgere alle associazioni professionali, ai centri di formazione, agli enti con un portafoglio di iniziative internazionali e ai PCO, cioè tutti quei soggetti che si occupano di organizzare eventi e fornire servizi linguistici.[20]

Infine, nella fase pratica di raccolta vera e propria (attraverso la registrazione e la presa di note sul campo) vanno affrontati diversi aspetti di natura tecnica, i quali assumeranno caratteristiche specifiche in ogni singola circostanza, a seconda del tipo di impianto acustico e di video-trasmissione utilizzato, nonché della conformazione fisica della sala in cui si trovano i vari partecipanti. L'ideale sarebbe poter contare su di una stretta collaborazione tra chi si occupa della ricerca e chi organizza il convegno, in modo da includere le necessità di raccolta dati nella prassi organizzativa, cioè già a partire dalla fase precongressuale (Palazzi 1999, pp. 54-55; Shalom 2002). Probabilmente, questo comporterebbe una migliore assistenza tecnica, nonché la possibilità di includere nelle condizioni di ingaggio degli interpreti la richiesta di consenso alla registrazione e all'uso dei dati per scopi accademici, chiarendo ovviamente i termini di riservatezza e anonimato del caso. Lo stesso varrebbe per tutti gli altri partecipanti, i quali sarebbero informati fin da subito della raccolta dei dati. Lo strumento con cui poter concretizzare l'accesso ai dati e sancire la collaborazione dei soggetti interessati è il consenso informato.

[20] A tal riguardo, una testimonianza interessante è stata fornita da Viel (comunicazione personale) nella fase di raccolta dei dati per la sua tesi di laurea. Nel 2010 Viel ha contattato i 31 membri dell'associazione interMED (interpreti specializzati in ambito medico), ricevendo ben cinque risposte positive. Sarebbero stati gli stessi interpreti e sconsigliarle di contattare gli organizzatori anticipatamente in merito alla registrazione, per il timore che questi non avrebbero mostrato altrettanta disponibilità. Pertanto, in tutte le occasioni di raccolta (tre o quattro) la richiesta veniva anticipata solamente al PCO, i cui responsabili, pur confermando anche la loro disponibilità, si sono sempre rifiutati di interloquire sul momento con gli organizzatori dei vari convegni al fine di evitare problemi. Questa esperienza dimostra inequivocabilmente che la richiesta di collaborazione (cioè il consenso informato) va inoltrata anticipatamente a tutti i soggetti coinvolti, ponendo la massima cura alla gestione delle pubbliche relazioni.

1.3.2.2 Consenso informato

Oltre a essere accessibili, è fondamentale che i dati da raccogliere siano liberi da restrizioni d'uso e distribuzione dovute al copyright o ad altri tipi di vincoli. Si tratta di una delle questioni più spinose che da sempre limitano la possibilità di studiare non solo l'interpretazione, ma anche altre forme di comunicazione.[21] Nello specifico, il consenso deve essere fornito non solo dagli organizzatori o dai responsabili dell'evento registrato, ma anche dai partecipanti (oratori) e dagli interpreti coinvolti. Come purtroppo è confermato in più di una esperienza di ricerca, tendenzialmente sarebbero proprio gli interpreti ad opporre maggiori resistenze, a causa del timore «of exposing mistakes or weaknesses in the translated material» (Armstrong 1997, p. 154).

> Questa resistenza sembra derivare dal timore che l'analisi del testo interpretato porti a un giudizio dell'analista sulle capacità professionali dell'interprete. È un preconcetto dovuto a una certa tradizione di studi: se il testo d'arrivo viene confrontato col testo di partenza per vedere quanto è stato reso del senso originario e con quale accuratezza (si veda la lunga tradizione degli studi sugli errori in interpretazione), allora può essere percepita una componente valutativa, a giusta ragione sgradita.
> (Zorzi 2004, p. 75)

Come giustamente osserva ancora Zorzi, la situazione sembra ora in una fase di cambiamento grazie a una maggiore consapevolezza del valore della ricerca, nonché alla presenza di *practisearchers*, ovvero di "interpreti professionisti-ricercatori" – studiosi che presentano un profilo "ibrido" essendo professionisti con una specifica formazione nel campo dell'interpretazione di conferenza e al contempo ricercatori. Il primo tipo di formazione è, in particolare, uno degli ingredienti chiave non solo per cogliere al meglio i tanti aspetti che sono pertinenti all'attività di ricerca in questione, ma soprattutto per essere inseriti nella rete di professionisti attivi in una parte del mercato, i quali potrebbero offrire più facilmente la loro collaborazione in virtù del rapporto professionale o anche di amicizia già esistente con il *practisearcher*.

Per capire meglio i motivi che potrebbero aver alimentato questa antica diffidenza degli interpreti professionisti nei confronti della ricerca, va sottolineato che la traduzione prodotta dagli interpreti è innanzitutto parte di un

[21] Ricordiamo, per esempio, come gli autori di diversi lavori sulla comunicazione parlata presentati in occasione del IV convegno LREC (tenuto nel 2004 a Lisbona) ammettessero in sede di dibattito di non essere riusciti ad ottenere il consenso alla diffusione dei dati studiati per motivi legati alla proprietà intellettuale. Riguardo alla questione dei diritti d'autore per i testi scritti e pubblicati, si vedano Bowker & Pearson (2002, p. 59) e McEnery et al. (2006, pp. 77-79 A9).

servizio fornito nell'istantaneità del momento in cui si realizza la comunicazione. Di conseguenza, non appena ci si distacca dal contesto di realizzazione e fruizione immediata, si innesca inevitabilmente un processo di allontanamento, il quale se da una parte è utile all'osservazione dei dati, dall'altra può implicare una sorta di snaturamento dei dati stessi.[22] Tale rilevanza del TA entro i confini dell'immediatezza in cui viene prodotto dagli interpreti e fruito dagli utenti è espressa, per esempio, nella formulazione della clausola di non responsabilità presente nel sito del Parlamento europeo, precisamente nella Biblioteca Multimediale da cui si ha accesso alle registrazioni di tutte le sedute plenarie a partire da aprile 2006:

> Le registrazioni delle interpretazioni non rientrano nella documentazione ufficiale.
> L'interpretazione dei dibattiti ha il solo scopo di facilitare la comprensione dei partecipanti alla riunione. Le registrazioni, dunque, non rientrano nella documentazione ufficiale.
> Solo i discorsi originali trascritti o tradotti costituiscono documenti ufficiali.
> Qualsiasi altro utilizzo delle registrazioni delle interpretazioni per scopi diversi da quelli sopra enunciati è vietato ed è subordinato al rilascio di un'autorizzazione esplicita e specifica da parte del Parlamento europeo.
> (Biblioteca Multimediale PE)

L'uso delle registrazioni dei dibattiti tenuti durante le sedute plenarie del PE e delle relative interpretazioni è consentito per scopi accademici (non a fini di lucro) e per attività di divulgazione sui temi dell'Unione europea. Pur trattandosi di un tipo specifico di situazione comunicativa, con caratteristiche peculiari riguardo alle modalità di realizzazione, agli eventi linguistici, alla lingua e ai partecipanti coinvolti (de Manuel 2003b, p. 30; Bendazzoli 2010), è innegabile lo straordinario valore di questi materiali, soprattutto a fronte del grado di accessibilità e delle possibilità d'uso consentite.
La situazione è ben diversa nel contesto dei convegni internazionali organizzati nel mercato locale italiano, così come in molti altri paesi. Qui il consenso informato è uno strumento indispensabile per la tutela di tutti coloro che risultano coinvolti nella ricerca: dal ricercatore stesso ai partecipanti (compresi gli interpreti), i cui interventi sono registrati e diventano oggetto di studio.

[22] Questo comporta chiare conseguenze a livello metodologico e, soprattutto, se ne dovrebbe sempre tenere conto in sede di discussione dei risultati in modo da relativizzarli adeguatamente: «Targeting a given text in a specific situation (within a particular type of meeting involving participants with different sociocultural backgrounds) upon the users of interpreting is not an exceptional case; at least in conceptual terms, it should be seen as the rule» (Pöchhacker 1992, p. 218).

Più in generale, come segnalano Milroy & Gordon (2003, pp. 79-87) si tratta di una vera e propria questione etica. La preparazione di un documento con cui informare i partecipanti alla ricerca è un'operazione generalmente richiesta per tutti i tipi di studi che prevedono il coinvolgimento di esseri umani, con motivazioni di fondo più o meno pressanti (gli studi di sociolinguistica non sono certamente come gli studi clinici). Ciò che deve emergere chiaramente dal documento redatto per il consenso informato è che «Subjects must voluntarily agree to participate in the research and must know what their participation entails» (*ibid*., p. 79). Attingendo da Neuman (1997, p. 450), i due autori spiegano che il consenso si ottiene nella prassi attraverso la firma di una dichiarazione scritta, la quale dovrebbe contenere in linea di massima le seguenti indicazioni:

a) descrizione del progetto e degli obiettivi
b) descrizione delle procedure utilizzate
c) anonimato e riservatezza
d) conferma che la partecipazione è volontaria e che può essere ritirata in qualsiasi momento
e) informazioni di contatto dei ricercatori e dell'ente che li sponsorizza.

Abbiamo ripreso alcuni esempi riportati da Johnstone (2000, pp. 44-47) in riferimento alla ricerca sociolinguistica:

Figura 1.3 Esempio di modello di consenso informato nella ricerca sociolinguistica (Johnstone 2000, p. 44).

Sample 4.1 Certification of Informed Consent

Professors Barbara Johnstone and Judith Mattson Bean are carrying out a study of Texas women's speech. As part of the study, they will audiotape interviews with nine women, of whom I will be one. Drs. Johnstone and Bean have discussed the project with me. The project will result in publications in academic journals and, it is hoped, eventually in a book in which my speech may be quoted, described, and analyzed. If I have provided other materials relating to myself, my work, or my speech, or if other materials are publicly available, these materials may also be quoted, described, and analyzed. The intent of the study is to describe Texas women's speech, not to evaluate it.

By signing this form, I certify that Drs. Johnstone and Bean's research project has been satisfactorily explained to me and that I consent to participate in it in the ways described above.

Name: _____

Signature: _____

Date: _____

Figura 1.4 Esempio di modello di consenso informato nella ricerca sociolinguistica (Johnstone 2000, pp. 45-47).

Sample 4.2 Certification of Informed Consent

My name is Mary Bucholtz. I am a graduate student at the University of California at Berkeley. I would like you to take part in my research on the language of friendship among teenagers and young adults. I want to find out how people talk about friendship, and how they talk to their friends.

If you and your parent(s) or guardian(s) agree that you may participate in the research, I will meet you at [your high school] during your lunch hour or at another time and place that we agree on. There I will talk to you for about thirty minutes about your friendships at your school and what you think about friendship generally. I will record our interview on a tape-recorder and I may ask you to wear a small clip-on microphone during the recording.

After our interview, I may want to talk to you in more detail, with some of your friends. I will ask all of you about what you like to do together, what you like about each other, and how you became friends. I may also ask to meet with you and your friends outside of school to get an idea of how you talk to each other when you are relaxing together. I may ask to tape-record some of these conversations; if I do, I may again ask you to wear a clip-on microphone.

During the interview(s), I may ask you who you like and dislike, and who you think others like and dislike. I may also find out other private things about your life if I later meet with you and your friends. You may not want others to find out these things, so I will protect your privacy in every way I can:

- I will not let other people listen to the tapes or read transcripts or notes based on our conversation(s) unless you give your permission for them to do so. I will keep the tapes and writings locked up in a safe place in my home, where only I can get to them. I will not tell your friends, teachers, or family members what you say during our conversation(s).
- In writing and talking about my study, I will never use your name or any other names that could give away your identity. You will choose a name that I will use to refer to you. I will also use invented names for other people in your life so that no one can use that information to figure out who you are. Papers with your real name on them—this form and a note to myself reminding me what your invented name is—will be locked up in my home, separately from each other and from the tape-recordings and transcripts. In addition, the note reminding me what your invented name is will have only your first name on it.
- If at any point in the study you want me to erase from the tape anything you've said, you have the right to tell me to do this, and I will erase it in your presence.

After this research is completed, I may save the tape and notes for use in future research. However, I will protect your privacy in the future in the same way that I will protect it during the study.

Your participation in this research is voluntary. You are free to refuse to take part and you may refuse to answer any questions or may stop taking part at any time. Whether or not you participate will have no effect on your standing at your school.

If you have any questions about the research, you may call me, Mary Bucholtz, at [phone number]. If you and your parent(s) or guardian(s) agree that you may take part in this research, please

return a signed copy of this form to me when we meet or send it to [address]. You should keep one copy of this form for your records.

Your signature on each part of the consent form indicates that you give permission for each part of the study. If you are under age 18, one of your parents or guardians must sign each part of the form as well. You may choose to sign only some parts of the form and not others.

1. I agree to take part in this research.

Student signature _____ Date _____
Parent/Guardian signature _____ Date _____

2. I give permission for the tape-recordings to be transcribed. I understand that the transcript will not include any names or details that will give away my identity.

Student signature _____ Date _____
Parent/Guardian signature _____ Date _____

3. I give permission for the transcripts to be used for teaching purposes and for presentation of the research at conferences.

Student signature _____ Date _____
Parent/Guardian signature _____ Date _____

4. I give permission for parts of the tape-recordings to be used for teaching purposes and for presentation of the research at conferences.

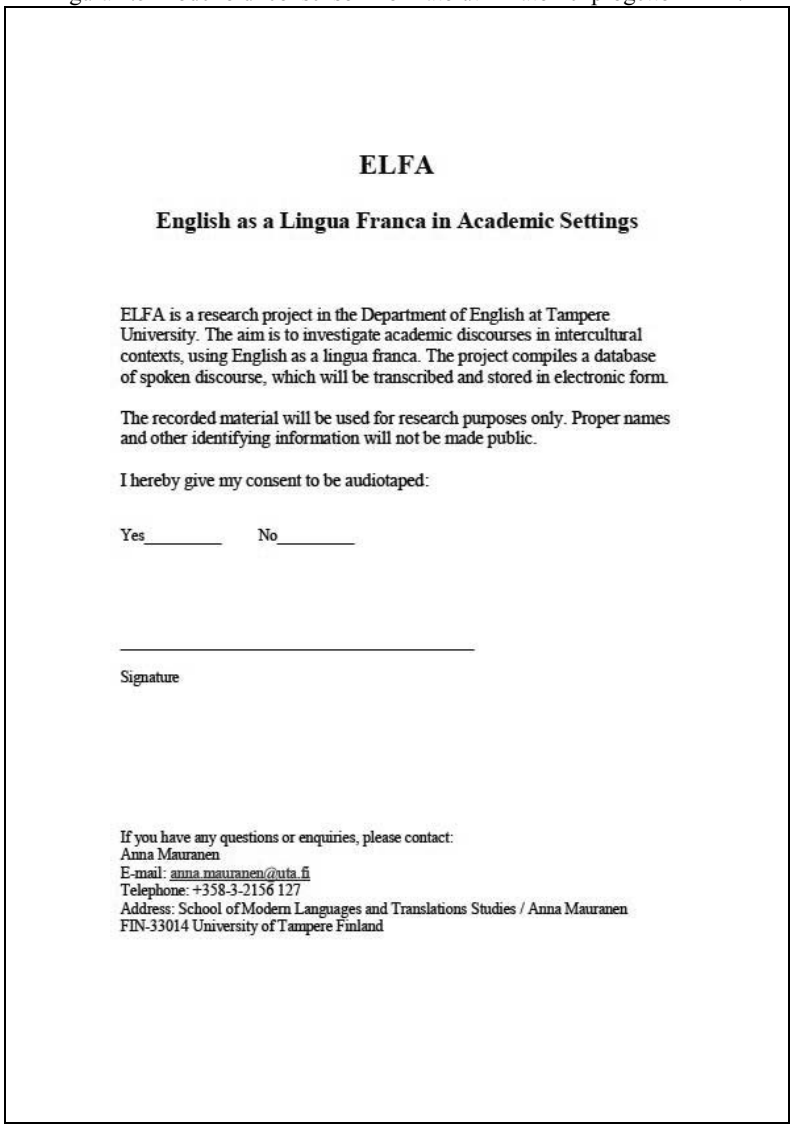
Student signature _____ Date _____
Parent/Guardian signature _____ Date _____

Data la natura prettamente linguistica delle ricerche di nostro interesse, sembrerebbe addirittura ammesso ottenere il consenso anche solo verbalmente: «written consent may not be required in every case; it may be acceptable to obtain consent verbally as long as subjects are properly informed» (Milroy & Gordon 2003, p. 80). Pare che esistano anche casi in cui il consenso informato non è strettamente necessario, come per gli studi che prendono dati di dominio pubblico o per i sondaggi svolti mantenendo l'anonimato delle persone. In altri casi, si cerca di favorire l'ottenimento del consenso indicando che nello studio è prevista la sostituzione dei nomi con pseudonimi, iniziali o numeri, a garanzia dell'anonimato. Un ulteriore aspetto da assicurare è il controllo attento sia degli utenti che possono accedere alle registrazioni e ai dati in generale, sia della possibile diffusione ad altri ricercatori, per uso a fini di ricerca e didattici (§1.3.6).

Infine, tra i tanti tipi di informazioni da inserire nel documento per il consenso informato, potrebbero essere aggiunti alcuni brevi accenni sul beneficio che scaturirebbe eventualmente dalla realizzazione della ricerca proposta, sulla base dell'idea che sia giusto restituire qualcosa alla comunità da cui si prendono i dati. Per esempio, gli studi sulla lingua di conferenza potrebbero fornire osservazioni utili su determinate condotte linguistiche o prassi comunicative, specialmente in riferimento alla gestione realizzata dagli interpreti e a come questa possa essere favorita e facilitata. In questo modo, insomma, il messaggio che verrebbe trasmesso è che si intende svolgere uno studio non solo "su" una certa comunità, ma anche "per" la comunità stessa.

Dopo aver messo a fuoco i punti essenziali che dovrebbero essere contenuti in un documento per il consenso informato, siamo andati alla ricerca di esempi già utilizzati precedentemente in altri studi di ambito traduttologico. Purtroppo, i modelli di consenso utilizzati da altri ricercatori non sono quasi mai resi accessibili, né sono solitamente divulgati nella letteratura assieme alla descrizione dei progetti o dei risultati ottenuti. Uno dei pochi esempi facilmente accessibili dalla Rete è il modello di consenso redatto per il progetto ELFA a cui abbiamo accennato prima:

Figura 1.5 Modello di consenso informato utilizzato nel progetto ELFA.



ELFA

English as a Lingua Franca in Academic Settings

ELFA is a research project in the Department of English at Tampere University. The aim is to investigate academic discourses in intercultural contexts, using English as a lingua franca. The project compiles a database of spoken discourse, which will be transcribed and stored in electronic form.

The recorded material will be used for research purposes only. Proper names and other identifying information will not be made public.

I hereby give my consent to be audiotaped:

Yes_____ No_____

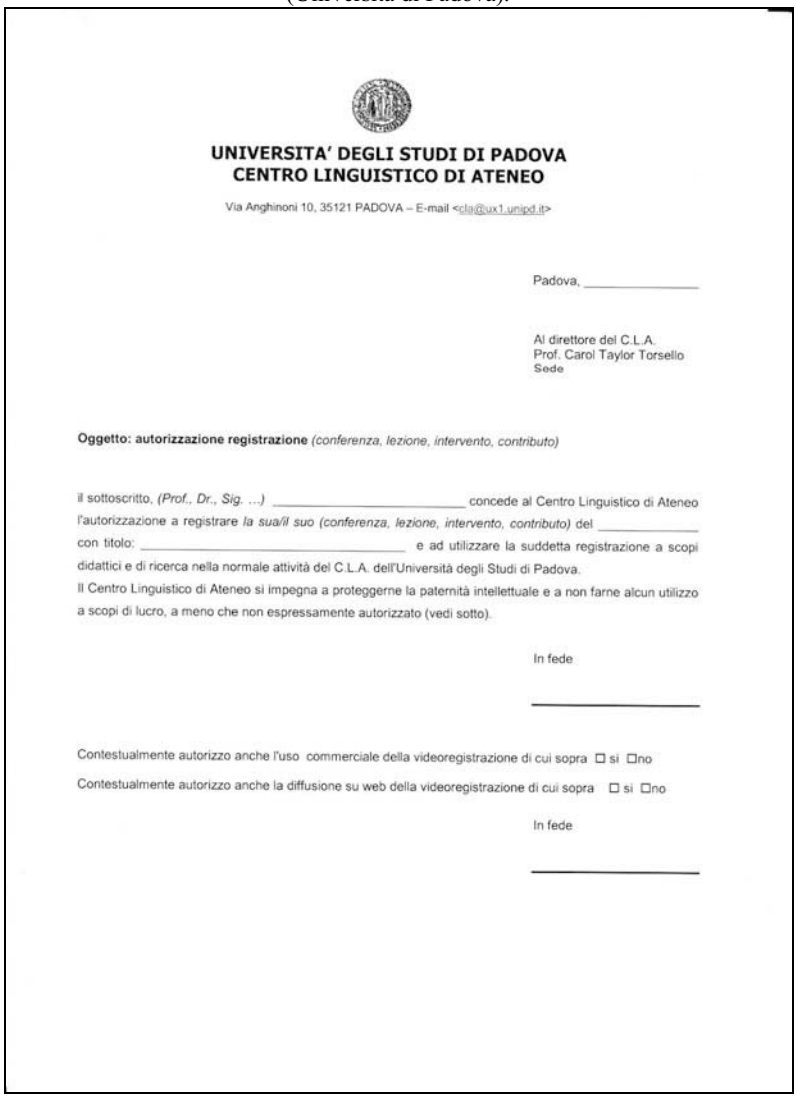
_____

Signature

If you have any questions or enquiries, please contact:
Anna Mauranen
E-mail: anna.mauranen@uta.fi
Telephone: +358-3-2156 127
Address: School of Modern Languages and Translations Studies / Anna Mauranen
FIN-33014 University of Tampere Finland

Un altro esempio trovato in Rete è il modello di consenso stilato dal Centro Linguistico di Ateneo dell'Università di Padova:

Figura 1.6 Modello di consenso informato utilizzato dal Centro Linguistico di Ateneo (Università di Padova).



UNIVERSITA' DEGLI STUDI DI PADOVA
CENTRO LINGUISTICO DI ATENEO

Via Anghinoni 10, 35121 PADOVA – E-mail <cla@ux1.unipd.it>

Padova, _____

Al direttore del C.L.A.
Prof. Carol Taylor Torsello
Sede

Oggetto: autorizzazione registrazione *(conferenza, lezione, intervento, contributo)*

il sottoscritto, *(Prof., Dr., Sig. ...)* _____ concede al Centro Linguistico di Ateneo l'autorizzazione a registrare *la sua/il suo (conferenza, lezione, intervento, contributo)* del _____ con titolo: _____ e ad utilizzare la suddetta registrazione a scopi didattici e di ricerca nella normale attività del C.L.A. dell'Università degli Studi di Padova.
Il Centro Linguistico di Ateneo si impegna a proteggerne la paternità intellettuale e a non farne alcun utilizzo a scopi di lucro, a meno che non espressamente autorizzato (vedi sotto).

In fede

_____

Contestualmente autorizzo anche l'uso commerciale della videoregistrazione di cui sopra ☐ si ☐no

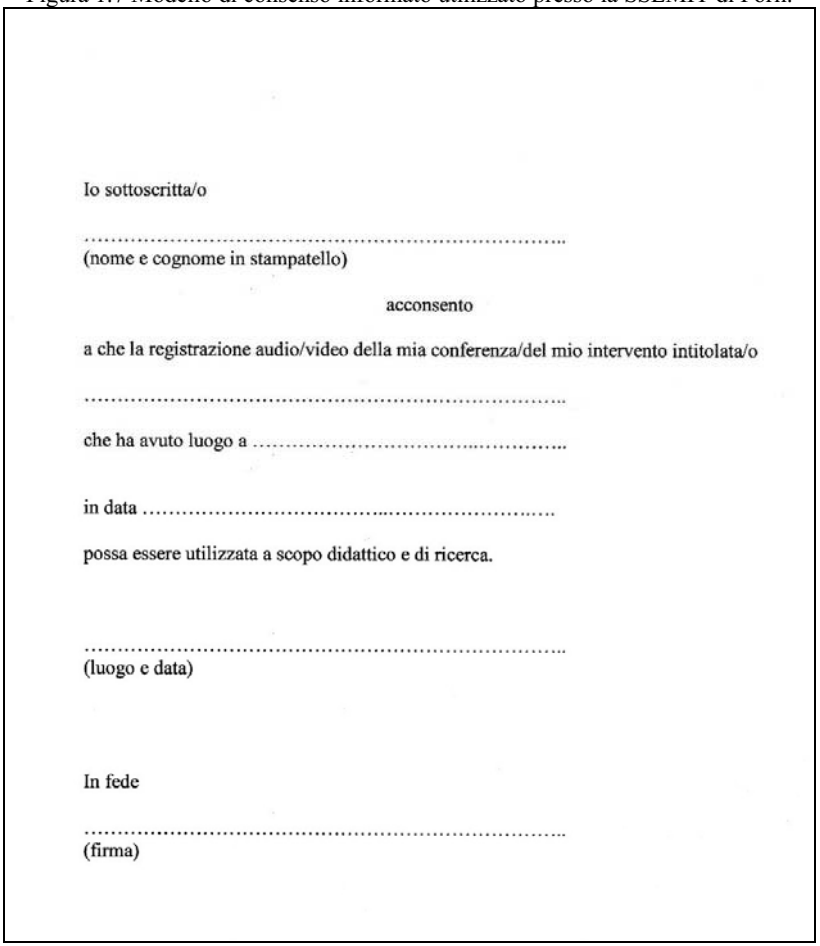
Contestualmente autorizzo anche la diffusione su web della videoregistrazione di cui sopra ☐ si ☐no

In fede

_____

Infine, due esempi di moduli per il consenso informato utilizzati specificatamente in ambito traduttologico sono il modello usato presso la SSLMIT di Forlì e il modello utilizzato da Raffaela Merlini nel raccogliere esempi di consecutiva breve in contesti reali:

Figura 1.7 Modello di consenso informato utilizzato presso la SSLMIT di Forlì.



Io sottoscritta/o

……..
(nome e cognome in stampatello)

acconsento

a che la registrazione audio/video della mia conferenza/del mio intervento intitolata/o

……..

che ha avuto luogo a ……..……..

in data ……..……..

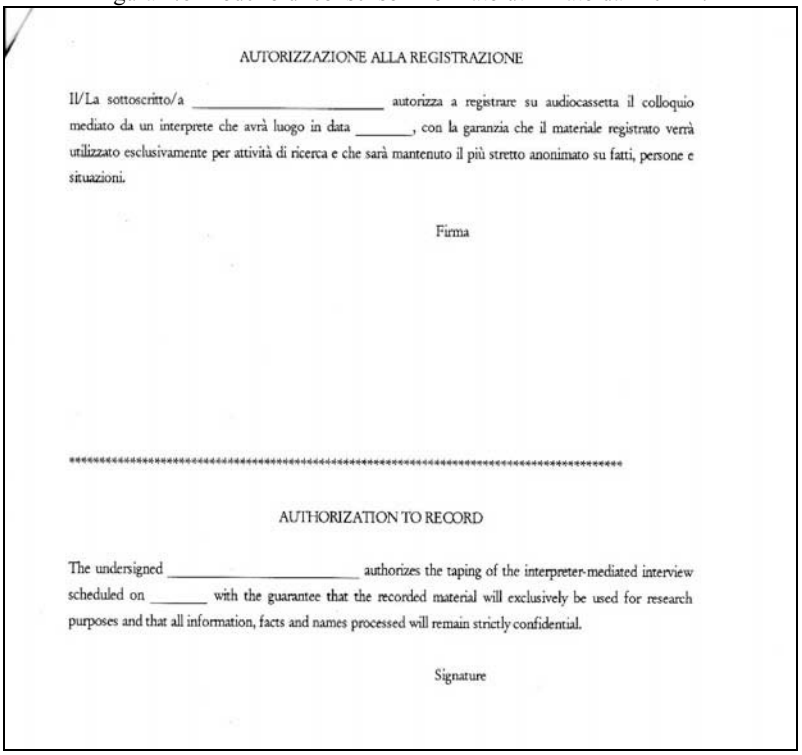
possa essere utilizzata a scopo didattico e di ricerca.

……..
(luogo e data)

In fede

……..
(firma)

Figura 1.8 Modello di consenso informato utilizzato da Merlini.



AUTORIZZAZIONE ALLA REGISTRAZIONE

Il/La sottoscritto/a _____ autorizza a registrare su audiocassetta il colloquio mediato da un interprete che avrà luogo in data _____, con la garanzia che il materiale registrato verrà utilizzato esclusivamente per attività di ricerca e che sarà mantenuto il più stretto anonimato su fatti, persone e situazioni.

Firma

****

AUTHORIZATION TO RECORD

The undersigned _____ authorizes the taping of the interpreter-mediated interview scheduled on _____ with the guarantee that the recorded material will exclusively be used for research purposes and that all information, facts and names processed will remain strictly confidential.

Signature

Come abbiamo affermato precedentemente, è piuttosto raro trovare in letteratura indicazioni esplicite su come le questioni della riservatezza dei dati e dell'ottenimento del consenso alla registrazione da parte degli interessati siano state affrontate nella metodologia di ricerca. In una delle rare eccezioni che abbiamo rilevato, è interessante notare come Wadensjö (1998, pp. 98-102) sia pienamente consapevole dei metodi della ricerca sociologica rispetto all'uso del consenso informato e delle pratiche di raccolta dei dati, probabilmente perché opera in un ambito istituzionale con dati decisamente più "sensibili" rispetto al contesto dei convegni internazionali come eventi aperti al pubblico. Questo è evidente nella necessità di puntualizzare che tra le norme di riservatezza garantite ai suoi interlocutori viene data enfasi alla possibilità di sospendere in qualsiasi momento la disponibilità a farsi registrare, nonché all'anonimato. Oltre a questo, particolare considerazione è riservata al fatto di «not to make the interpreters feel stressed about being observed and recorded; not to make them experience this as a collection of the 'interpreter's errors'» (*ibid*., p. 98).

Anche Diriker (2004, p. 56) affronta questa parte metodologica della sua ricerca e spiega come il fatto di conoscere personalmente gli organizzatori dell'evento da lei analizzato, così come di essere collega degli interpreti in servizio abbia reso questa fase di raccolta priva di particolari ostacoli e difficoltà.[23] Tuttavia, non è specificato se fossero state prese misure di alcun tipo con altri partecipanti, ovvero gli oratori e il pubblico, nei confronti dei quali probabilmente era stato ritenuto sufficiente ottenere il beneplacito degli organizzatori.

In un contesto sempre di conferenza-convegno, ma privo di alcun servizio di interpretazione, Räisänen adotta la seguente metodologia:

> The recordings of the presentations were done by myself using a small portable DAT recorder. The presentations were also recorded centrally by a technician. Before the opening of the conference, a committee member announced that the joint session was being recorded. Consent to use the recorded material was obtained from each individual speaker after his/her presentation. Obtaining permission beforehand might have affected the production of their texts, and asking their permission just before their presentations could have made them nervous and affected their performance. (Räisänen 1999, p. 21)

Vedremo fino a che punto la nostra esperienza nel reperire i materiali per il progetto DIRSI coincide più o meno con quella documentata nei due contributi illustrati sopra (§4.2.2). Dato il nostro interesse attorno alla modalità di raccolta dei dati negli Studi sull'interpretazione su base empirica e sul campo, nel corso degli ultimi anni abbiamo spesso posto direttamente la questione a vari studiosi che presentavano i loro dati nel corso di convegni e giornate di studio. Per esempio, nella creazione del suo corpus elettronico di interpretazione simultanea e consecutiva, Meyer (2008, si veda §2.3) non ha nemmeno avvertito la necessità di procedere alla stesura di un modello di consenso informato (Meyer, comunicazione personale). Questa decisione è stata motivata dal fatto che tutti i soggetti coinvolti, l'associazione organizzatrice della serie di conferenze studiate, la relatrice (si trattava sempre della stessa persona) e gli interpreti, avevano già collaborato in precedenza con l'istituto di afferenza di Meyer e i rapporti erano tali per cui non ci sarebbe stato nessun tipo di problema legato al consenso e alla registrazione. Nello specifico, Meyer e il suo gruppo di ricerca hanno avuto la possibilità di accedere a tutti i dati necessari, comprese informazioni dettagliate sugli interpreti professionisti in servizio (formazione,

[23] Un'altra testimonianza simile è di Kurz (comunicazione personale), la quale è riuscita a ottenere la collaborazione di interpreti professionisti (Kurz 2002, 2003) proprio grazie alla sua appartenenza diretta alla comunità linguistica interessata (sia gli interpreti, sia i formatori).

esperienza, profilo linguistico), in quanto erano i principali sovvenzionatori dell'evento stesso. Gli interpreti hanno accordato il loro consenso oralmente e i loro nomi non appaiono all'interno della ricerca, pur mettendo a disposizione tutte le informazioni menzionate.

Ad ogni modo, sia nei casi simili alle testimonianze di Meyer e Diriker, sia nei contesti in cui è invece assolutamente necessario redigere un consenso e ottenere la firma dei soggetti interessati, si deve sottolineare l'esistenza di un passaggio estremamente delicato ancora più a monte, ovvero il contatto con i soggetti a cui il ricercatore si rivolge per presentare la richiesta di registrazione e perché tale richiesta sia successivamente inoltrata a tutti gli interessati. Il più delle volte, è in tale frangente che il ricercatore si gioca tutte le possibilità di accesso ai dati, pertanto è indispensabile non solo preparare con accuratezza il modello di richiesta di collaborazione e di consenso, ma anche individuare il contatto migliore e studiare le modalità più efficaci con cui avvicinarlo. Oltre che trovare conferma di questo nella nostra esperienza di ricerca, lo stesso Franz Pöchhacker (comunicazione personale) ha dichiarato di aver potuto accedere ai dati di cui ha riferito in occasione del convegno Critical Link 5[24] poiché poteva contare sulla collaborazione di una persona che già operava all'interno del contesto da dove sarebbero stati presi i dati; questo tipo di contatto diventa, utilizzando le parole di Pöchhacker, un *inside champion*. I dati raccolti in questo caso specifico sono di natura estremamente sensibile, in quanto provengono da un contesto giuridico in cui sono coinvolte persone con lo status di rifugiati. Conseguentemente, si può ritenere che l'accesso ai dati e il consenso al loro uso per scopi di ricerca abbiano avuto un esito felice quasi del tutto esclusivamente grazie al sostegno che solo il cosiddetto *inside champion* ha potuto garantire. Se è vero, come risulta anche dal nostro studio, che un meccanismo simile può essere riscontrato in molti altri contesti, è altrettanto vero che è fondamentale procedere con estrema cautela al coinvolgimento di potenziali *champion*, nonché garantire loro la dovuta tutela da possibili conseguenze negative. Inevitabilmente, un *inside champion* si trova esposto a critiche da parte di coloro che non hanno particolare interesse a collaborare con la ricerca, ma che si trovano coinvolti, loro malgrado, per motivi istituzionali. Tutto questo presuppone l'adozione delle misure ritenute necessarie affinché il grado di esposizione alle critiche del nostro *inside champion* sia mantenuto al minimo, per esempio garantendone l'anonimato e non eccedendo nella frequenza e nella quantità di richieste di collaborazione. Ogni situazione va valutata caso per caso.

Una volta assicurata l'accessibilità ai dati che ci interessa studiare, il passo successivo prevede la loro registrazione. Si tratta di un'operazione non

[24] Pöchhacker (2007), Pöchhacker & Kolb (2007); il convegno *Critical Link 5 – Quality in Interpreting: A Shared Responsibility* si è svolto dall'11 al 15 aprile 2007 a Parramatta – Sydney (Australia). Si veda inoltre Kolb & Pöchhacker (2008).

priva di difficoltà e per la quale è bene essere consapevoli di vari aspetti e accorgimenti.

1.3.2.3 Registrazione

Come per molte altre discipline che studiano la comunicazione parlata, una prassi consolidata per raccogliere i dati consiste nel registrare l'interazione che si vuole analizzare. Ovviamente, la registrazione video sarebbe da preferire al solo audio, in quanto essa può fornire un quadro più completo dell'evento in esame. Tuttavia, si tratterebbe comunque di una visione "limitata" perché ottenuta da una prospettiva particolare e soggettiva (l'obiettivo della telecamera nel punto in cui è collocata). Nel caso dei convegni, per esempio, a seconda del tipo di studio si dovrebbe posizionare la telecamera tra il pubblico (rivolta verso il tavolo dei conferenzieri), nella cabina per la simultanea o comunque vicino ad essa (così da catturare l'evento dalla prospettiva degli interpreti), o ancora presso il tavolo dei conferenzieri con l'obiettivo rivolto verso il pubblico. Sono tutte prospettive plausibili all'interno dello stessa situazione comunicativa e, al contempo, diverse tra loro. Inoltre, se già la possibilità di ottenere il consenso alla registrazione è ardua nel caso delle "meno invasive" registrazioni audio, la presenza di una telecamera potrebbe sollevare più resistenze di quante ne vengano poste normalmente. In realtà, l'ostacolo maggiore è forse da attribuire principalmente alla disponibilità dell'attrezzatura tecnica e della gestione dei dati video da parte dei ricercatori, i quali probabilmente trovano più semplice effettuare registrazioni audio e più agevole la gestione di dati in un formato più leggero. Il problema ovviamente non si pone nei casi in cui la registrazione sia curata dagli organizzatori (benché questo potrebbe comportare altre criticità, dovute ad eventuali errori commessi dai tecnici o dal mancato invio dei materiali a chi è interessato ad analizzarli).

Sono molte le considerazioni pratiche e metodologiche al riguardo. Procedendo con ordine e facendo riferimento alla registrazione dei due flussi comunicativi (TP e TA), approfondiremo le principali criticità relative ai seguenti aspetti: strumentazione tecnica, formati e programmi informatici, gestione dei dati.

1.3.2.3.1 Strumentazione tecnica

Il numero di canali audio attraverso cui è trasmessa la comunicazione in un evento mediato da interpreti dipende direttamente dalla modalità di traduzione coinvolta e dal numero di lingue utilizzate. Le modalità in differita prevedono generalmente una trasmissione su un canale unico (quello del TP), mentre con le

modalità di interpretazione in simultanea sono impiegati più canali sovrapposti contemporaneamente: il TP è trasmesso su un canale solitamente indicato con il termine *floor* anche nelle apparecchiature di ricetrasmissione; ogni TA è trasmesso invece su un canale individuale per ciascuna lingua. Nel caso di convegni con solo due lingue di lavoro, pertanto, i flussi comunicativi sono due, uno per il TP e uno per il TA, ma i canali di trasmissione sono tre: uno per il TP e due per il TA (uno per ciascuna delle due lingue di lavoro coinvolte). Al fine di catturare tutti i dati, ci si può avvicinare ad essi seguendo diverse strategie, in base a un approccio per così dire "esterno" o a un approccio "interno".

Il progresso tecnologico ha reso possibile la registrazione e la gestione di quantità di dati prima inimmaginabili. Come spiegato precedentemente (§1.3.2.1), si può accedere ai dati non solo "dall'interno" dell'evento in cui sono prodotti, ma talvolta anche "dall'esterno", come è il caso dei canali televisivi satellitari o della Rete (de Manuel 2003b, pp. 27-35). Se fino a non molto tempo fa i ricercatori erano soliti utilizzare registratori a batteria e supporti su nastro magnetico con una quantità spaventosamente in crescita di audiocassette[25] o videocassette,[26] oggi possono avvalersi di strumentazioni digitali, della tecnologia Internet e della trasmissione dei dati a banda larga. I vantaggi sono ovvi e non è certo necessario elencarli in questa sede; basti pensare alla possibilità di registrare i dati senza soluzione di continuità (non bisogna "girare la cassetta")[27] e alla facilità con cui si possono ottenere diverse copie.

Quando si accede ai dati "dall'interno", emerge immediatamente la questione del coinvolgimento diretto del ricercatore nella loro raccolta. A questo proposito, Gile (1998) distingue tra *interactive* e *non-interactive obervational research*, dove la prima

[25] Si veda l'esperienza di Kalina (1994, p. 226). Dopo essere riuscita ad ottenere il consenso alla registrazione in un convegno trilingue con servizio di interpretazione simultanea, ha lasciato la sala del convegno portando con sé oltre 20 audiocassette.

[26] Ricordiamo come ancora nel 2004, nell'ambito della realizzazione del corpus EPIC (§3) fu necessario registrare tutti i materiali su videocassetta VHS da quattro diverse postazioni televisive, ciascuna munita di decoder satellitare (Monti et al. 2005). I registratori DVD e prodotti simili di acquisizione video non analogici erano all'epoca al loro debutto nel mercato italiano. Per questo, fu necessario digitalizzare in seguito tutte le registrazioni (140 videocassette VHS) utilizzando uno specifico programma informatico di acquisizione video e audio in un computer dedicato esclusivamente a tale scopo. Sebbene una copia dei dati su supporto magnetico costituisca un ulteriore backup, addirittura l'esperienza relativa a quello che è considerato il primo esempio di *spoken corpus* aveva mostrato che la digitalizzazione «is extremely time-consuming, and takes up a lot of storage» (Knowles 1993, p. 107). Sorprendentemente, lo stesso vale anche per i corpora di lingua scritta, laddove siano considerati testi di cui non è disponibile una versione in formato digitale. In questi casi, si rende necessaria la scansione ottica e la conversione dei caratteri così ottenuti, oppure l'uso del riconoscimento vocale per velocizzare la riscrittura del testo al computer (Bowker 2002, pp. 22-42).

[27] Si vedano le trascrizioni raccolte in appendice allo studio di Galli (1988/1989), nelle quali sono riportate interruzioni nel testo dovute proprio alla fine della cassetta (un esempio è a pagina 127).

> […] involves participation of the investigator in the process under study and/or questionnaires or interviews, and thus entails the risk of interference by the researcher and/or a significant influence of the research procedure on the phenomenon under study, or the risk of substantial interference from the subjects' personal perception, interpreting and reporting facts.
> Gile (1998, p. 74)

Pur rappresentando un limite, la ricerca interattiva condotta da *practisearchers* (la maggior parte dei quali registra anche le proprie prestazioni) può contribuire a sensibilizzare maggiormente la popolazione oggetto di studio perché vi sia sempre più collaborazione. Inoltre, i dati raccolti possono poi essere studiati anche da altri ricercatori, superando così del tutto il limite posto dalla autoanalisi. Kalina (2005, p. 35) osserva come questo tipo di ricerca sia stata forse l'unica finora a consentire di svolgere studi su campioni di dati più consistenti ed estrapolati da situazioni reali, a fronte delle difficoltà e dei limiti che abbiamo discusso. «The few lager audio recorded corpus data that exist mostly stem from conferences where some of those engaged in research were directly involved (Pöchhacker's corpus for his hypertext approach, 1994[b], and Kalina's Würzburg corpus, 1998, which has also been used by Setton, 1999»).[28] Un altro esempio è lo studio di Vik-Tuovinen (2000, p. 18) in cui sono analizzate le registrazioni dei commenti e dei dialoghi scambiati tra gli interpreti in cabina mentre non stanno fornendo il servizio. Anche in questo caso, pur rappresentando un limite, il coinvolgimento in prima persona del ricercatore anche come interprete per l'evento comunicativo in questione è stato con buona probabilità la chiave che gli ha consentito di accedere ai dati (nonché di avere il consenso per la registrazione).[29]

Riprendendo il riferimento al *practisearcher* e alla presenza del ricercatore all'interno dell'evento comunicativo che si intende studiare, è necessario essere consapevoli di quanto la presenza di un osservatore (interno o

[28] Un'opinione del tutto simile è espressa anche dallo stesso Setton (s.d.): «Poechhacker (1994) recorded a 3-day conference on small businesses in Vienna, where his role as recruiter and coordinator of the interpreting team gave him a unique overview of the whole conference as a multilingual interpreted event».

[29] Diversa è ovviamente la situazione negli studi sperimentali, dove gli unici partecipanti ai quali è necessario chiedere il consenso sono gli interpreti e comunque solo i soggetti in esame nello studio. In questi casi, il coinvolgimento di interpreti professionisti può essere ottenuto più facilmente attraverso un ingaggio vero e proprio dietro compenso, oppure grazie alla sensibilità dei professionisti che sono anche impegnati in attività di formazione. Ad esempio, si veda come è stato gestito questo aspetto anche attraverso l'utilizzo di interviste retrospettive negli studi di Ivanova (2000, pp. 33-34) e Chang & Schallert (2007, p. 171). Ad ogni modo, il coinvolgimento diretto del ricercatore all'interno della situazione comunicativa in cui si stanno raccogliendo i dati sembrerebbe essere una prassi diffusa anche negli studi sul parlato spontaneo. Si veda, a tal proposito, quanto descritto da Moneglia (2005, pp. 216-217) sui corpora di LABLITA.

esterno) possa influenzare lo svolgersi dell'evento e, nel nostro caso specifico, la resa degli interpreti e l'esposizione dei conferenzieri. Alcuni potrebbero proporre di effettuare le registrazioni di nascosto, assumendosi la responsabilità di tutto ciò che è implicato in prassi indicate come *surreptitious recording* o *candid recording* (Milroy & Gordon 2003, pp. 81-83). Tali prassi possono essere accettate in alcuni casi, spiegando che è stata effettuata una registrazione solo successivamente, ma dovrebbero essere evitate nei confronti di gruppi grandi di soggetti con i quali non sono già attivi rapporti di conoscenza. Riteniamo che nel nostro ambito si debba obbligatoriamente informare i partecipanti della registrazione, anche a tutela di studi futuri che altrimenti troverebbero ancor più difficoltà nel reperimento dei dati. Questo è in linea con il principio generale espresso da Labov (1984, p. 52 citato in Milroy & Gordon 2003, p. 83) secondo cui è indispensabile «To avoid any act that would be embarrassing to explain if it became a public issue».

Quali sono le strategie possibili per registrare il flusso comunicativo dei partecipanti primari, ovvero di coloro che emettono il TP? L'uso di un registratore nella sala del convegno comporterebbe la registrazione di tutti i rumori ambientali, con un livello qualitativo insufficiente se si intendesse utilizzare i materiali raccolti per scopi didattici. In alternativa, l'audio del *floor* può essere acquisito direttamente in un computer portatile, collegandolo all'impianto di amplificazione utilizzato in sala. Il personale tecnico in servizio ha qui ovviamente un ruolo fondamentale, poiché è il soggetto a cui si deve rivolgere la richiesta di collegare il proprio computer all'impianto di amplificazione. In alcuni casi fortunati, può darsi che allo stesso tecnico sia già stato richiesto un servizio di registrazione (anche video) dei lavori del convegno da parte degli organizzatori. Questo potrebbe valere anche per l'uscita audio degli interpreti, magari con una acquisizione su doppia pista. Nonostante questa opzione sembri essere la più vantaggiosa (e comoda), in realtà lo è fino a un certo punto. Infatti, risulta rischioso (e allo stesso tempo poco corretto) affidare totalmente al tecnico di sala la responsabilità di registrare i dati e di attivare, quindi, in tempo utile la registrazione per non perdere i primi secondi di inizio dei lavori. Nell'esperienza già menzionata di Diriker troviamo sconcertanti testimonianze dei problemi che possono insorgere nel caso in cui il ricercatore non abbia modo di controllare la registrazione (Diriker 2004, pp. 56-57). La collaborazione con il tecnico è certamente indispensabile, ma non si deve correre il rischio di intralciare la sua funzione e di aggiungersi ai problemi che spesso si trova costretto a risolvere (con una sala piena di persone che attendono il tocco magico che rimetta in sesto qualsiasi tipo di apparecchiatura mal funzionante). Inoltre, nel malaugurato caso in cui gli interpreti si dimenticassero di accendere il microfono o di sintonizzarsi sul canale di uscita corretto in tempo utile, potrebbero perdersi porzioni preziose del TA.

Quali sono le alternative a disposizione per registrate l'altro flusso comunicativo, nel quale si ha la traduzione del *floor*, ovvero il TA trasmesso dagli interpreti? La risposta più immediata è l'uso di registratori digitali di dimensioni ridotte, più comodi da utilizzare dei vecchi registratori a cassetta e, soprattutto, "meno invasivi" se utilizzati nell'ambiente ristretto della cabina. A questo proposito, un'accortezza essenziale riguarda l'uso delle batterie che forniscono alimentazione elettrica a questo tipo di apparecchiature. Al fine di prevenire la perdita di dati preziosi, è consigliabile utilizzare batterie nuove all'inizio di ogni convegno, se non addirittura sostituirle prima della ripresa dei lavori nel caso il convegno sia suddiviso in una parte mattutina e una parte pomeridiana. Pur non essendo molto rispettosa dell'ambiente, questa strategia dovrebbe essere adottata sempre, anche quando sembra che vi sia ancora una quantità di carica sufficiente a disposizione. Nel suo studio, sempre Diriker (2004, p. 58) lamenta una perdita di dati proprio per questo motivo.

Un altro aspetto da segnalare riguardo all'uso di registratori in cabina è che in questo modo potrebbe essere registrato anche il TP, percepibile come "rumore di fondo" a volume più o meno elevato a seconda del livello di insonorizzazione della cabina stessa. Sicuramente, lo stesso effetto si otterrebbe se si utilizzasse il registratore all'esterno della cabina, appoggiando un auricolare per la ricezione del TA degli interpreti al registratore stesso (Firenze, comunicazione personale). Tuttavia, quest'ultimo stratagemma garantisce un livello di qualità della registrazione eccessivamente vulnerabile; la qualità non migliora nemmeno nel caso in cui si colleghi uno dei ricevitori forniti al pubblico (assieme agli auricolari per sentire gli interpreti) con un cavo audio a un registratore o a un computer per l'acquisizione digitale. In questa eventualità, oltretutto, si dovrebbe avere l'accortezza di sintonizzarsi sempre sul canale corretto di uscita degli interpreti, in quanto il canale effettivo del TA cambierebbe a seconda della lingua dell'oratore nel TP.

L'impiego di un registratore digitale non è l'unica possibilità di registrazione della resa degli interpreti. Al di là dell'eventualità sopra menzionata in cui il tecnico di sala si faccia carico di registrare TP e TA su doppia pista, si potrebbe posizionare un microfono apposito all'interno della cabina e collegarlo a un computer portatile posto all'esterno, gestito totalmente dal ricercatore. In questo caso, il ricercatore riuscirebbe difficilmente a gestire la registrazione se fosse anche coinvolto come interprete nel convegno in questione. Ad ogni modo, un primo vantaggio in questa opzione è dato dalla qualità della registrazione (l'acquisizione su computer da microfono è decisamente migliore di quanto si possa ottenere con un registratore in cabina). Inoltre, non si pone il limite delle batterie per l'alimentazione del registratore. Infine, l'assenza del registratore nella cabina eliminerebbe l'eventuale condizionamento che tale apparecchiatura potrebbe provocare al lavoro degli interpreti. Per contro, l'ultima soluzione proposta sarebbe particolarmente

dispendiosa in termini di equipaggiamento tecnologico che il *practisearcher* sarebbe costretto a portarsi al seguito.

1.3.2.3.2 Formati, programmi e archiviazione dei dati

Grazie all'avvento del digitale, non è più necessario rifarsi ai soli supporti magnetici per registrare i dati audio e video in qualsiasi contesto.[30] Oltre alle apparecchiature appositamente pensate per questo scopo, come i (mini) registratori, le videocamere e le attrezzature presenti in una sala di regia, esistono molte altre applicazioni utilizzabili direttamente dal computer, e che consentono di realizzare acquisizioni audio e video collegando un microfono o una webcam. In genere, questi programmi sono dotati di funzioni che permettono anche di editare il file acquisito, migliorandone le caratteristiche, e di salvarlo in molteplici formati. Alcuni esempi sono Audacity, Cool Edit Pro e WaveLab per i file audio; Pinnacle Studio e Movie Maker per i file video.

I formati a disposizione in ciascun programma per salvare un file sono numerosi. I fattori principali da considerare per orientarsi nella scelta dei formati da utilizzare sono la gestione agevole dei dati raccolti e la possibilità di scambiarli con altri ricercatori o di utilizzarli con più applicazioni. A seconda della mole di dati inclusi in ogni singola registrazione, i file potrebbero acquisire un "peso" più o meno consistente. Tuttavia, la scelta di formati "leggeri" (come .MP3) comporterebbe la perdita di informazioni importanti (per esempio nel caso di analisi spettrografiche). Oltre a questo, è buona prassi fare uso di formati che siano quanto più compatibili con i sistemi di funzionamento dei riproduttori audio e video, evitando cioè di salvare i dati in formati che risultano leggibili solo da una specifica applicazione. Un formato per i file audio che risponde ai requisiti esposti è il formato .WAV. Per i file video la situazione è forse più delicata, poiché le dimensioni dei file potrebbero essere tali da dover utilizzare supporti molto capienti, quali hard disk esterni e spazi server. Anche in questo caso, i progressi tecnologici rendono la questione sempre meno difficile da gestire. Tuttavia, la sfida non tocca soltanto i dati registrati, ma anche i documenti contenenti le trascrizioni di tali dati. Senza entrare ora in dettaglio sul tema delle trascrizioni (si veda la sezione successiva), vale la pena puntualizzare che la scelta dei formati è un aspetto cruciale anche per i documenti di testo che diventeranno la base del corpus. Questo è parimenti valido in altri ambiti di ricerca, come esemplificato nella seguente testimonianza (Thompson 2005):

[30] Significativo è il riferimento all'uso dei floppy disk nel primo progetto di *spoken corpus* (Knowles 1993, p. 107).

> It is usually hoped that corpora will be made available for other researchers to use, and in this case it is necessary to create a corpus that is in a suitable format for interchange of the resource. There are also closely related issues to do with preservation of the resource […]. I recently requested a copy of an Italian corpus called PIXI [corpus may be ordered from OTA] from the Oxford Text Archive. The files for the corpus are in WordPerfect format, and I opened them using both a text editor, and WordPerfect 9. As the 'Readme' file informs me, there are some characters in the files that are specific to WordPerfect and which do not convert to ANSI. Some of these characters were used in this corpus to mark overlap junctures.
>
> The version I see shows:
>
> <S C><p> $$Li avevo gia?presi, esatto.%% Poi pero? $ne avevo ordinati -% <S A><p> $Allora aspetti che guardiamo% se e?rimasto:
>
> This shows how the particularities of word-processors and the character sets that they use create problems in interchange between different programmes, even between different versions of the same word-processing package.

Generalmente, per le trascrizioni è consigliabile salvare i file in formato "testo" puro (.TXT), evitando così di includere informazioni "nascoste" sulla formattazione dei caratteri utilizzati (come succede invece con MS Word). In questo modo, si dovrebbero evitare eventuali problemi di leggibilità da parte di diversi programmi, specialmente per quel che riguarda gli accenti, gli apostrofi e gli eventuali segni diacritici. Tra le applicazioni disponibili per gestire file di testo, oltre al comune Blocco note presente in tutti i PC, TextPad presenta numerose funzioni, assai utili anche per la ricerca di particolari parole o caratteri all'interno di un file o di gruppi di file.

Prima ancora di produrre le trascrizioni, già dall'inizio della fase di raccolta dei dati attraverso la registrazione si presenta la necessità di gestirli e catalogarli adeguatamente. Si crea, infatti, fin da subito un vero e proprio archivio multimediale, nel quale troveranno spazio anche i file di testo che si ottengono con le trascrizioni dei TP e dei TA. Per questo motivo, è importante impostare un sistema di gestione efficace, con il quale si riesca non solo a trovare velocemente i dati necessari, ma anche ad estrarre informazioni utili alla descrizione del campione raccolto. Tale sistema potrebbe basarsi su fogli di lavoro Excel, oppure su archivi Access o simili. Dall'altra parte, l'immagazzinamento delle registrazioni comporta l'uso di supporti sufficientemente capienti e affidabili, come hard disk esterni, server, nonché di strumenti per il effettuare il backup e preservare in sicurezza di tutti i dati, conservandone la massima qualità.

1.3.3 Trascrizione

Dopo aver completato le prime due tappe nella creazione di un corpus di interpretazione (*corpus design* e *data collection)*, la fase successiva prevede che i dati raccolti siano trascritti. Questo passaggio essenziale dal parlato allo scritto solleva non poche sfide metodologiche, con le quali già molti altri *Translation scholars* si sono confrontati (Armstrong 1997, Shlesinger 1998, Meyer 1998, Falbo 2005). Più in generale, la questione è da lungo tempo oggetto di dibattito e riflessione tra tutti coloro che studiano la comunicazione parlata, indipendentemente dal fatto che essa sia mediata da un interprete o meno (Bazzanella 1994, Blanche-Benveniste 2005). Oltre che sul piano delle considerazioni di tipo teorico, il tema della trascrizione merita di essere approfondito anche dal lato pratico, esaminando gli strumenti e le alternative di realizzazione di ciò che rappresenta la base di uno *spoken* corpus. A tal fine, i principi avanzati da Edwards (1993a) rappresentano un utile punto di partenza per discutere entrambe le dimensioni. Il primo principio, *category design*, è il fulcro della riflessione teorica sulla funzione che la trascrizione avrebbe nel rappresentare i dati per consentirne l'analisi; dall'altra parte, il secondo e il terzo principio, *readablity* e *computational tractability*, permeano anche la dimensione pratica dell'attività di trascrizione, poiché strettamente legati all'uso concreto degli strumenti di analisi.

1.3.3.1 Considerazioni teoriche

A causa delle difficoltà insite nell'analizzare la comunicazione parlata spontanea, Nencioni (1989) spiega come molti studi si siano addirittura basati su testi di parlato "simulato", attingendo cioè da testi drammatici o altri testi di natura scritta nei quali era "riprodotta" la lingua parlata:

> […] data la difficoltà del registrare il parlato in condizioni di spontaneità totale e poi di trascriverlo adeguatamente, inferenze valide sui modi parlati possono essere tratte anche da testi scritti, purché simulino competentemente il parlato e vengano analizzati con cautela.
> (Nencioni 1989, p. 241)

Fortunatamente, la situazione nel tempo è andata radicalmente cambiando e sarebbe oggi impensabile svolgere uno studio sulla comunicazione parlata, tanto più se mediata da interpreti, sulla base di testi interamente simulati.

Pur constatando che gli ostacoli al reperimento e alla registrazione di materiali autentici vanno riducendosi sempre più, la trascrizione dei dati orali resta un'attività che richiede ancora uno sforzo considerevole in termini di tempo, energia e attenzione. "«El principal inconveniente es la cantidad de tiempo que consumen las transcripciones del material grabado […]»" (de Manuel 2003b, p. 58) è una constatazione assai ricorrente tra chi si occupa di studi sul parlato. Specialmente l'investimento in termini di tempo, in effetti, è tanto più vero per coloro che si occupano di interpretazione, in quanto a un TP da trascrivere si accompagna sempre un TA all'incirca della stessa durata e in un'altra lingua (una puntualizzazione non banale nel valutare le dimensioni di un corpus di interpretazione).[31] Anche Kalina (1994), nonostante l'entusiasmo di essere riuscita a reperire una considerevole quantità di dati registrati, spiega come si sia subito resa conto della necessità di farsi assistere nell'attività di trascrizione al fine di poter svolgere il suo studio in tempi ragionevoli. Non a caso, i grandi progetti di corpora orali monolingui (§1.2.1) hanno generalmente alle spalle una folta squadra di trascrittori professionisti o di ricercatori con a disposizione risorse cospicue.

Per quanto l'aggettivo *time-consuming* sia ancora spesso usato tra coloro che affrontano il compito della trascrizione, la disponibilità di registrazioni in formato digitale ha portato sostanziali vantaggi. Tra questi, vi è la possibilità di usare applicazioni che consentono di riascoltare automaticamente il brano che si sta trascrivendo, o che permettono di controllare la registrazione attraverso i tasti della tastiera del computer, potendo rallentare o mettere in pausa il file audio/video più agevolmente (per esempio SoundScriber, VoiceWalker e Transana). Allo stesso modo, anche i programmi di riconoscimento vocale sono diventati un ausilio significativo per velocizzare i tempi di scrittura, e quindi anche di trascrizione (alcuni esempi sono IBM ViaVoice e Dragon NaturallySpeaking). Con questi ultimi programmi è possibile infatti dettare i testi da trascrivere anche mentre si ascoltano, senza interruzioni di sorta, eseguendo quindi un esercizio di *shadowing* (ripetere a voce alta ciò che si sta ascoltando in cuffia seguendo il ritmo del testo originale).

Al di là degli strumenti con cui produrre le trascrizioni, resta da esaminare la fondamentale questione di ciò che implica la trasformazione di un testo orale in un testo in forma scritta, una trasformazione indispensabile ai fini dell'analisi. La trascrizione serve, appunto, a fissare la natura effimera del parlato, è «il mezzo con cui i significati creati dalla particolare lingua vengono "esternati" (espressi) in forma visiva anziché parlata» (Halliday 1992, p. 82), così da poter

[31] Stando a Hofland (2003), «In general, 10-15 hours of work are required in order to transcribe and manually time align one hour of speech». L'investimento di tempo risulta, infatti, tanto maggiore quanto più è sofisticato il sistema di annotazione applicato alle trascrizioni.

"sentire con gli occhi" e "vedere con le orecchie", in una sorta di ribaltamento dei canali di percezione. Tuttavia, non si deve cadere nella tentazione di considerare il testo trascritto al pari di un testo scritto a pieno titolo, poiché quest'ultimo è normalmente prodotto e fruito in condizioni totalmente diverse, essendo il risultato finale di un processo di produzione nel quale sono sì comprese anche eventuali correzioni, cancellazioni e così via, ma in modo non manifesto. Per contro, nel parlato non è possibile celare le operazioni di "elaborazione redazionale" del testo prodotto, le quali risultano manifeste a tutti i partecipanti alla situazione comunicativa (e andrebbero pertanto trascritte).[32] Volendo dunque paragonare direttamente il testo tra-scritto al testo scritto, sarebbe opportuno considerarlo come un testo scritto nel suo formato di bozza, contenente tutti i segnali di riparazione, riformulazione e modifica presenti nel corso della sua produzione, quali elementi rivelatori dei meccanismi che sottostanno al funzionamento della comunicazione (Blanche-Benveniste 2005, pp. 21-28; Halliday 1992, p. 80).

Avendo chiarito il senso dell'uso del mezzo scritto per rappresentare la comunicazione parlata, è evidente che la trascrizione va considerata una forma di rappresentazione statica della lingua, ed è quindi uno strumento di ausilio all'analisi dei dati audio/video. Come tale, presenta tutta una serie di limiti derivanti da più fattori, tra cui le possibilità di impiego dello spazio tipografico;[33] l'elaborazione seriale in unità discrete di elementi che provengono da un flusso composto da più livelli sovrapposti; la diversa articolazione sintattica tra parlato e scritto; la gestione di eventuali elementi linguistici non standard; l'indeterminatezza dei fenomeni paralinguistici e non verbali; le specificità del sistema di scrittura selezionato; l'eventuale uso di strumenti di analisi automatica (Orletti & Testa 1991, pp. 245-250).

Uno dei sistemi di trascrizione più conosciuti e largamente utilizzato negli studi sulla comunicazione parlata (specialmente per la conversazione spontanea) prende il nome di "Sistema jeffersoniano", poiché fu sviluppato da Gail Jefferson per poi essere utilizzato, in particolare, nell'ambito dell'Analisi della conversazione (Psathas & Anderson 1990, Atkinson & Heritage 1999).[34] Esso comprende una varietà di notazioni particolari e di simboli, con cui raccogliere ed esprimere in forma scritta le caratteristiche più salienti rilevabili all'interno di una interazione orale (per esempio, le sovrapposizioni, l'organizzazione in turni conversazionali, gli intervalli all'interno dello stesso enunciato, la contiguità tra

[32] Questo è vero ad eccezione dei prodotti multimediali che sono il risultato di un processo di post-produzione.

[33] Sull'influenza che la conformazione "fisica" della trascrizione può avere sulla lettura dei dati si vedano Ochs (1999) e Edwards (1995).

[34] Per una panoramica su altri sistemi di trascrizione, con particolare riferimento alla conversazione spontanea, si vedano O'Connel & Kowal (1994, 1999, 2009) e Edwards (1993a); per alcuni esempi di sistemi notazionali si vedano Blanche-Benveniste (2005, pp. 55-64) e Du Bois et al. (1993); sui diversi ambiti, non solo accademici, in cui viene fatto uso di diversi tipi di trascrizione si veda Mack (2006).

due diversi enunciati o *latching*, gli allungamenti vocalici, i troncamenti di parola o frammenti, l'enfasi, l'altezza, il volume, il respiro e così via). Questo sistema non ha certo la pretesa di trasporre in forma scritta tutte le dimensioni dell'oralità, ma si presenta come un chiaro esempio di «analytic interpretation and selection» (Psathas & Anderson 1990, p. 75) dei dati orali – una interpretazione e una selezione motivate innanzitutto dalla presenza stessa dell'analista.[35]

> Ciò che deve essere chiaro a proposito di una trascrizione di questo tipo è il fatto che prima che il lettore si trovi con questi 'dati', si è già avuto un certo lavoro di interpretazione da parte dell'analista. […] In questo processo di creazione della versione scritta del testo orale egli fa appello ai modi convenzionali di interpretazione che egli pensa siano condivisi da altri parlanti della stessa lingua.
> (Brown & Yule 1986, p. 23)

Sarebbe a tutti gli effetti impossibile, per non dire pretenzioso e inutile, tentare di includere tutti i tratti dell'oralità possibili e immaginabili in una trascrizione.

> Incluso usando un sistema de transcripción que trate de reflejar ciertas peculiaridades del discurso oral, especialmente del conversacional, la fijación gráfica del habla implica transformar un proceso dinámico en un producto textual estático, implica atribuir secuencialidad a lo simultáneo (proxémico-gestual, paraverbal, suprasegmental), e inevitablemente conlleva perder de vista muchos de los elementos comunicativos presentes en el habla.
> (Recalde & Vázquez Rosa 2009, p. 60)

Il «carattere selettivo della trascrizione» (Orletti & Testa 1991, p. 250) è strettamente legato all'obiettivo di ciascuna ricerca e gioca un ruolo fondamentale nel preservare la leggibilità dei dati trascritti. Per questo motivo, si ritiene importante non delegare il compito della trascrizione a soggetti esterni alla ricerca (Psathas & Anderson 1990, p. 77); allo stesso modo, il sistema utilizzato dovrebbe essere sufficientemente flessibile da poter essere compatibile

[35] Non si tratta solo di un processo di interpretazione; la produzione stessa del testo trascritto è altrettanto influenzata dalla capacità di svolgere questa operazione da parte di chi trascrive, poiché potrebbero essere commessi degli errori. La tendenza a normalizzare eventuali disfluenze e ad anticipare il testo, con conseguenti errori di trascrizione, è stata provata anche sperimentalmente (Chiari 2006). Pertanto, trascrivere rimane un'attività da non sottovalutare in termini anche di attenzione e capacità se si vogliono evitare «slips of the ear – slips which show themselves as errors of transcription» (O'Connel & Kowal 1999, p. 107).

con altri obiettivi. In definitiva, l'impostazione del sistema notazionale in uno studio dovrebbe essere orientata dai seguenti fattori:

- comprensibilità vs. specializzazione (grado di inclusione dei fenomeni ritenuti rilevanti)
- attendibilità (uso della notazione per indicare i fenomeni rilevanti)
- leggibilità (sia da parte dell'analista, sia da parte della macchina)
- consistenza interna (evitare l'uso di uno stesso simbolo per più fenomeni)
- flessibilità (adattarsi alle esigenze della ricerca)
- trasversalità (rispettare l'uso di notazioni consolidato anche in altri studi)
- riproducibilità (compatibilità con l'elaborazione automatica).

A questo punto, il tipo di trascrizione che si intende produrre dipenderà dal tipo di studio che ci si propone di realizzare. Una prima distinzione di massima vede due tipi di trascrizioni, chiamati *broad transcription* e *narrow transcription* (Du Bois et al. 1993). Il primo tipo «includes the most basic transcription information: the words and who they are spoken by, the division of the stream of speech into turns and intonation units, the truncation of intonation units and words, intonation contours, medium and long pauses, laughter, and uncertain hearings or indecipherable words» (*ibid*., p. 46). A ben vedere, per quanto "*broad*" questo tipo di trascrizione contiene un buon numero di informazioni e tratti dell'oralità. Dall'altra parte, il secondo tipo includerebbe informazioni aggiuntive, quali «The notation of, among other things, accent, tone, prosodic lengthening, and breathing and other vocal noises» (*ibid*., p. 46).
Una seconda distinzione, più dettagliata, riprende le componenti essenziali del fenomeno in esame, la comunicazione parlata, alcune delle quali sono state approfondite proprio ai fini della trascrizione (O'Connel & Kowal 2009): la componente verbale, la componente prosodica, la componente paralinguistica e la componente extralinguistica.

Nella *componente verbale* si possono inquadrare quattro diversi tipi di trascrizione: la trascrizione ortografica standard, nella quale sono trascritte tutte le parole secondo la grafia standard della lingua utilizzata (normalizzando eventuali disfluenze); la trascrizione letterale, nella quale sono invece riprodotte esattamente le parole così come sono pronunciate, senza correzioni o adattamenti in caso di parole pronunciate in modo scorretto; la trascrizione secondo il metodo *eye dialect*, che implica «an effort to encode impressionistically all relevant sound categories» (*ibid*., p. 242), dando conto, quindi, di un maggior grado di deviazione rilevabile nella produzione dei suoni;

la trascrizione fonetica, basata sulle categorie fonetiche ed espressa utilizzando i caratteri dell'alfabeto fonetico internazionale.[36]

La *componente prosodica* include alcuni tratti non verbali, quali l'altezza, la durata e il volume dei suoni prodotti. Si tratta di caratteristiche piuttosto difficili da annotare in modo sistematico. Lo stesso Goffman (1981, p. 175) ha rimarcato la persistenza delle difficoltà con cui «[…] these paralinguistic markers can be satisfactorily identified, let alone transcribed». Questo vale anche per i tratti appartenenti alla *componente paralinguistica*, come le risate, il respiro, il pianto e altri tipi di vocalizzazioni.

> Se l'analista normalizza il testo orale in base alle convenzioni della forma scritta, le parole acquisiscono una formalità e una specificità che travisano necessariamente la forma parlata.
> I problemi che riguardano la rappresentazione segmentale delle parole pronunciate diventano insignificanti di fronte a quelli riguardanti la rappresentazione dei tratti sovrasegmentali (i particolari dell'intonazione e del ritmo). Non disponiamo di convenzioni standardizzate per rappresentare le caratteristiche paralinguistiche dell'enunciato, comprese nel termine 'qualità della voce' […].
> (Brown & Yule 1986, p. 22)

Infine, la *componente extralinguistica* concerne l'infinità di elementi che attorniano ciò che è compreso nelle precedenti componenti: essi vanno dai movimenti del corpo, ai fattori ambientali e così via.[37]

A questo punto, è opportuno precisare nuovamente che «One of the important features of a transcript is that it should not have too much information» (Ochs 1999, p. 168). Questo diventa ancora più valido se ci si prefigge di trascrivere un campione di grandi dimensioni, specialmente nei progetti di ricerca con l'obiettivo di realizzare un nuovo corpus (ovvero un D.I.Y. corpus). Come suggerito da Armstrong (1997) la trascrizione dei dati orali nei CIS dovrebbe inizialmente limitarsi a un livello minimo di rappresentazione del parlato. «This would include the words uttered (both complete and incomplete), pauses and filler sounds and basic segmentation,

[36] La scelta di un particolare tipo di trascrizione rispetto ad altri può influenzare notevolmente la fattibilità o meno di certi tipi di analisi. La testimonianza di Timarová (2005) ne è un esempio particolarmente interessante. Al fine di conteggiare il numero di sillabe contenute nel corpus dal lei studiato (in lingua ceca, §2.2), aveva dapprima trascritto alcune abbreviazioni per esteso, esattamente come erano pronunciate dai soggetti (l'esempio da lei riportato è "USA", trascritto come "u es a"). Successivamente, al fine di calcolare la densità lessicale «transcriptions had to be adjusted from the previous analysis (*u es a* back to *USA* to be counted as one word, etc.)» (*ibid.*, p. 68).

[37] Per un esempio di repertorio di simboli con particolare attenzione anche alla componente non verbale si veda Ochs (1999, pp. 175-181).

typically in terms of sentences» (*ibid*., p. 158); analogamente, Shlesinger (1998) consiglia di attenersi a un livello di trascrizione limitato alle caratteristiche più facilmente rappresentabili e, soprattutto, di utilizzare un sistema di codifica basato su convenzioni condivise. Rispetto all'orientamento convergente di queste due studiose, Bersani Berselli (2004, p. 65) avanza un dubbio più che legittimo: «il punto è se la rimozione dei tratti dell'oralità – che, è bene dirlo, semplificherebbe di gran lunga il successivo lavoro di annotazione del corpus – cancelli una o più dimensioni importanti di variazione». L'osservazione di Bersani Berselli va confermata per entrambe le opzioni: l'esclusione di un solo tratto dell'oralità dalle convenzioni di trascrizione di un testo escluderebbe automaticamente la considerazione immediata di tale tratto ai fini dell'analisi e, potenzialmente, offrirebbe una versione più limitata della possibile variazione[38] (si dovrebbe, in ogni caso, ricorrere sempre al dato orale/video). Dall'altra parte, la semplificazione (auspicabile) del lavoro di annotazione diverrebbe altrettanto vera, così come lo sarebbe l'accelerazione dell'opera di trascrizione stessa. È questo un punto cruciale nella realizzazione di un corpus di lingua parlata: per quanto in un metodo di trascrizione con convenzioni ridotte al minimo (come quelle adottate in EPIC e DIRSI-C) possano risultare assenti molte caratteristiche ritenute fondamentali, queste possono sempre essere aggiunte in un secondo momento, avvalendosi di strumenti e metodi scientifici,[39] e non tanto basandosi sulla mera percezione del trascrittore. Inoltre, potendo sfruttare la base di trascrizione già realizzata da altri che, consapevolmente, hanno prediletto la quantità alla "qualità" dei *token* contenuti nel corpus, l'aggiunta anche manuale di un nuovo livello di annotazione dovrebbe risultare meno ostica e garantire un livello di precisione e correttezza ottimali.

Quale che sia l'approccio adottato, la condizione imprescindibile è che vi sia sempre la possibilità di risalire ai dati audio/video (cioè alle registrazioni),[40]

---

[38] A questo proposito, le osservazioni formulate da O'Connel & Kowal (1994) dopo aver esaminato come un particolare fenomeno, quale l'aspirazione, è stato riportato nelle trascrizioni di numerosi studi sono alquanto rivelatrici. Infatti, non solo hanno rilevato una variabilità incredibile nelle modalità di notazione per indicare il fenomeno (distinto, in alcuni casi, tra espirazione e inspirazione), ma hanno anche potuto constatare che in fase di discussione non vi è mai la benché minima considerazione del fenomeno annotato, nel senso che i risultati non sono mai posti in relazione con tale fenomeno.

[39] Alcuni esempi di software che consentono di allineare il testo trascritto alla traccia audio o al filmato video corrispondente sono illustrati più avanti (§1.3.5.1).

[40] Si vedano le osservazioni di O'Connel & Kowal (1999, pp. 112-113) sul carattere ridondante, spesso immotivato, di molte trascrizioni, nelle quali la rappresentazione pseudoscientifica di alcuni tratti dell'oralità potrebbe essere risparmiata, rimandando alla registrazione vera e propria. Una posizione leggermente diversa è espressa da Leech (1997a, p. 4): «For a written corpus, the text itself is the data (in the etymological sense **data** are 'givens'), and the annotations are superimposed on it. For a spoken corpus, the recording is 'given', and it can also be maintained that a bare verbatim transcript of 'what was said' is itself a kind of 'secondary given', that is, a written record without any addition of less reliable, less clearly-definable, information.[…] Beyond these givens it is difficult to go without implicitly taking up some descriptive or interpretative stance towards the data». Tuttavia, si noti che anche in questo caso il dato trascritto si differenzia dal dato registrato in quanto "*secondary given*" e per il fatto che, a detta dello

potendo contare, per lo meno, su di un archivio costruito con una logica *user-friendly*, oppure, ancora meglio, effettuando l'allineamento tra le trascrizioni e le corrispondenti tracce audio/video (§1.3.5.1).

Prima di occuparci dell'annotazione, ovvero di esaminare i modi in cui le caratteristiche di variazione dell'oralità (e non solo) selezionate possono essere effettivamente inserite all'interno delle trascrizioni in un corpus elettronico, restano da affrontare alcune questioni di natura pratica, evidenziate dagli altri due principi menzionati in apertura di questa sezione (Edwards 1993a). Essi sono *readablity* e *computational tractability*, ovvero la distribuzione spaziale del testo trascritto, cioè il modo concreto in cui è visualizzato, e la sua leggibilità da parte di programmi informatici.

1.3.3.2 Considerazioni pratiche

Il principio di leggibilità della trascrizione, o *readablity*, proposto da Edwards (1993a) comprende due elementi, quali *visual prominence* e *spatial arrangement* (Edwards 1995). Essi sono particolarmente rilevanti nella rappresentazione del testo trascritto, soprattutto per quel che riguarda i turni di parola tra i partecipanti all'evento comunicativo che si vuole analizzare.[41] A seconda del tipo di ricerca e del formato interazionale da rappresentare, le trascrizioni possono essere strutturate secondo un'impostazione verticale, a colonna o a spartito. Nel formato verticale, i turni di ogni singolo partecipante sono posti in sequenza, uno sotto l'altro, garantendo così una certa equità nella percezione del tipo di apporto alla comunicazione da attribuire a ciascun partecipante; per contro, distribuendo i turni in colonne attigue, si tenderebbe a offrire una rappresentazione più forte di chi occupa la colonna a sinistra (questo sarebbe vero nelle culture che seguono un ordine di scrittura e lettura da sinistra a destra); infine, l'impostazione a spartito richiama la rappresentazione che si ha delle note in un pentagramma, pertanto parrebbe essere «the best of the three for display of densely overlapping events, including verbal and non verbal, if used with specialized software» (Edwards 1995, p. 27).

---

stesso Leech, «In rendering speech in written or electronic form […] a transcriber must necessarily *interpret* the discourse in the course of representing it» (*ibid*., p. 3). Schmidt (2009, p. 153) si spinge oltre nel puntualizzare che nei corpora orali i *primary data* sono costituiti dall'interazione stessa, la cui registrazione sarebbe da considerare come *secondary data*. Ne consegue che la trascrizione della registrazione andrebbe vista alla stregua di *tertiary data*, sottolineando in questo modo l'elevato tasso di *scientific modelling* a cui sono sottoposti i dati trascritti e annotati, usati in sede di analisi.

[41] Le stesse regole ortografiche della lingua scritta hanno inevitabilmente un impatto sulla rappresentazione dell'oralità a livello delle singole parole, oltre che a livello della struttura interazionale. Si pensi al caso dei nomi composti (per esempio *user-friendly* vs. *user friendly*), rappresentabili con o senza trattino e, pertanto, circoscrivibili come una o più parole individuali (Knowles 1993, p. 15).

Uno dei formati verticali più noti tra chi si occupa di comunicazione parlata (spontanea) è il formato CHAT, inizialmente sviluppato nell'ambito del progetto CHILDES per lo studio del linguaggio infantile (MacWhinney 1997, 2000). Esso è stato impiegato in molti altri progetti, tra cui l'edizione italiana del progetto CHILDES (Bortolini & Pizzuto 1997) e il progetto C-ORAL-ROM (Cresti & Moneglia 2005). Come illustrato da Moneglia & Cresti (1997), questo formato è stato pensato appositamente per gestire al meglio la rappresentazione dell'interazione dialogica e la sua annotazione, sulla base di tre tipi di righe: righe testa (*header*), righe di testo (la trascrizione divisa verticalmente in turni dialogici o battute che contengono in sequenza orizzontale gli enunciati di un parlante) e righe dipendenti (per le annotazioni). Un altro sistema particolarmente efficace per la trascrizione di scambi comunicativi dialogici o conversazionali è il cosiddetto sistema HIAT (*Halbinterpretative Arbeitstranskriptionen*, Ehlich 1993), basato sempre su una visualizzazione a spartito, con più livelli, in base all'idea per cui «one can consider simultaneous speech of several speakers at a time as a complex acoustic event, similar to the multitude of musical notes in a concerto» (*ibid.*, p. 131). In questo modo, diventa agevole la gestione della sincronicità di più fenomeni e più livelli di annotazione all'interno della trascrizione. Questo sistema è stato inizialmente implementato con un software apposito per l'inserimento dei dati chiamato syncWRITER, ma è pure compatibile con il software Exmaralda (§1.3.5.1), utilizzato anche nella realizzazione di studi sull'interpretazione simultanea e consecutiva (Meyer 1998, 2000, 2008; §2.3).

Nel caso delle trascrizioni di TP e TA, la validità maggiore o minore di uno o dell'altro formato dipenderebbe direttamente dalla modalità di interpretazione considerata (in differita o in simultanea), oltre che dall'obiettivo della ricerca. Setton (2002, p. 34) descrive due tipologie di visualizzazione con particolare riferimento alle modalità in simultanea (ma non esclusivamente), ossia una visualizzazione interlineare sincronizzata e una visualizzazione tabulare. Nella prima tipologia di visualizzazione (*synchronised interlinear transcript*), il TP e il TA occupano due righe separate e sovrapposte, così da rispecchiarne la sincronicità di emissione. Similmente, altre righe possono essere aggiunte per fornire la traduzione letterale di uno dei due flussi comunicativi o ulteriori livelli di annotazione. Nella seconda tipologia di visualizzazione (*tabular presentations*), il TP e il TA appaiono affiancati e distribuiti in colonne separate, una contenente il TP e una o più colonne per ogni TA, a seconda del numero di lingue diverse in cui è tradotto il TP, o del numero di TA nella stessa lingua ottenuti da più interpreti per lo stesso TP.[42] A queste

[42] Al fine di produrre e gestire al meglio i diversi formati di trascrizione, nel tempo sono state sviluppate applicazioni apposite, con le quali è possibile gestire i file multimediali (audio o video) unitamente ai file

due diverse opzioni di leggibilità delle trascrizioni, Setton ne aggiunge un'altra chiamata *fluent version* o versione "scorrevole" (revisionata) in quanto «punctuated, and with speech errors and hesitations eliminated» (Setton 2002, p. 35), di modo che il testo trascritto risulti più *reader-friendly*, cioè accessibile all'analista che volesse usufruirne con una normale lettura e senza eccessive difficoltà. Ovviamente, tale scelta ha parecchie implicazioni sull'uso effettivo, per non dire legittimo, di questa tipologia di trascrizione.

Infine, oltre a determinare il sistema notazionale con le convenzioni di trascrizione, nonché l'impostazione "grafica" più opportuna per visualizzare i dati, non ci si può esimere dal considerare anche il terzo principio messo a fuoco precedentemente (*computational tractability*), cioè la possibilità non solo di elaborare al computer il testo trascritto per poterlo poi recuperare più agevolmente, ma anche di analizzarlo con l'ausilio di applicazioni informatiche e programmi di linguistica computazionale. Questo è un ingrediente fondamentale se si intende applicare pienamente il *corpus-based approach* allo studio della comunicazione parlata e, in particolare, dell'interpretazione. Tuttavia, a livello pratico esso ha un impatto determinante sull'impostazione dei testi trascritti e delle relative annotazioni (§1.3.4), poiché le scelte effettuate dall'analista si troverebbero contese tra i due poli contrapposti della *machine-readability* e della *user/annotator-friendliness* (Chafe 1995, Cook 1995).[43]

In effetti, la piena applicazione del paradigma che è al centro del presente lavoro «remain[s] dependent upon a sound theory and practice of transcription whose principles are unaffected by the technology of storage, retrieval and analysis» (Cook 1995, p. 36). Affinché il computer sia in grado di elaborare le informazioni che sono incluse in ogni singola trascrizione (riconducibili alle componenti fondamentali illustrate prima), le stesse informazioni devono essere esplicitate, devono cioè essere espresse al pari del testo trascritto. L'aggiunta o l'esplicitazione delle informazioni rilevanti in un corpus si ottiene attraverso procedure di annotazione e codifica, le quali possono essere svolte in modo automatico, semi-automatico o manualmente.

---

di testo. Tra le più note ricordiamo Transana e Exmaralda, le quali consentono di produrre e visualizzare le trascrizioni secondo il modello tabulare o a spartito rispettivamente (si veda la sezione successiva per un approfondimento). Altri software che consentono di gestire la registrazione e, eventualmente, di inserire etichette temporali utili all'allineamento testo-suono (§1.3.5.1) sono VoiceWalker (Du Bois 2006b) e SoundWriter (Du Bois 2006a).

[43] O'Connel & Kowal (1999, p. 116) spiegano che «various readerships dictate radically different transcripts», in quanto l'analista, il lettore "esterno" e il computer presentano esigenze alquanto diverse.

1.3.4 Codifica e annotazione

La quarta tappa nella creazione di un corpus elettronico di interpretazione è strettamente connessa alla precedente tappa della trascrizione. La codifica e l'annotazione sono, infatti, operazioni con cui "arricchire" di informazioni i testi trascritti, secondo modalità particolari e tali da consentirne il successivo recupero semiautomatico attraverso il computer. Per questo motivo, le scelte operative su come rappresentare in forma scritta i dati orali (in altre parole, trascrivere le registrazioni) hanno un impatto diretto anche sulle possibili modalità di inserimento di dette informazioni.

La codifica, o *markup*, riguarda la descrizione o esplicitazione della struttura di un determinato testo, mentre l'annotazione riguarda un livello più specifico all'interno dei testi, vale a dire i loro aspetti linguistici e pragmatici (McEnery et al. 2006, pp. 22-28; Bowker & Pearson 2002, pp. 75-91). In particolare, l'annotazione

> […] can be defined as the practice of adding **interpretative, linguistic** information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process: the linguistic symbols which are attached to, linked with, or interspersed with the electronic **representation** of the language material itself.
> (Leech 1997a, p. 2)

Si notino nella citazione i tre diversi verbi utilizzati per riferirsi appunto a varie modalità di inserimento e aggiunta delle informazioni attraverso l'annotazione (*attached to*, *linked with*, *interspersed with*).

Solitamente, tra le informazioni sulla codifica, i dati metatestuali o contestuali sono raggruppati in una intestazione (*header*) all'inizio di ogni trascrizione, mentre le annotazioni pragmalinguistiche e metalinguistiche sono inserite direttamente all'interno del testo. In entrambi i casi, si presenta il problema della variabilità, nel senso che esistono diversi standard di codifica e di annotazione, le cui differenze renderebbero difficoltoso qualsiasi tentativo di paragone e scambio dei dati, così come ne impedirebbero l'interoperabilità (Leech et al. 1995, Leech 1997a, Kahrel et al. 1997). Si tratta di una questione da non sottovalutare, se non altro per il fatto che non solo un «annotated corpus is a more valuable resource than the original corpus», ma anche che la «corpus annotation tends to be an expensive and time consuming business» (Leech 1997a, p. 5).

A fronte di tutte queste criticità, un tentativo di standardizzazione è stato promosso dalla TEI (*Text Encoding Iniziative*, Burnard & Sperberg-McQueen

1994, Burnard 1995). La TEI è un consorzio che ha sviluppato un sistema di codifica standard per i testi elettronici e in formato digitale, al fine di armonizzare i diversi standard in uso ed evitare la proliferazione di diversi protocolli. Questo standard è stato creato inizialmente per i testi scritti, ma è stato poi esteso anche alla comunicazione parlata (Johansson 1995). Un'interessantissima applicazione di questo standard è stata inoltre effettuata in un corpus di trascrizioni tratte da eventi televisivi mediati da interpreti (Cencini 2002, Cencini & Aston 2002), adattando quindi la codifica TEI anche ai TA interpretati. Tale adattamento ha tenuto conto delle caratteristiche peculiari degli eventi mediali televisivi, nonché del ruolo che l'interprete assume nel fornire il proprio servizio in tali contesti e situazioni comunicative; a margine di questo lavoro pionieristico, è stata per di più proposta una versione di codifica alternativa, pensata per il contesto della conferenza. Di seguito riportiamo l'intero *header* e la parte iniziale della relativa trascrizione in cui abbiamo evidenziato gli elementi inseriti come attributi portatori di informazioni di varia natura (Cencini 2000, pp. 193 e ss.):

```
<?xml version="1.0"?>
<?xml-stylesheet href="TIC.css" type="text/css"?>
<!-- The TIC.dtd was compiled using the website
www.hcu.ox.ac.uk/TEI/newpizza.html;
the tagsets chosen are
base: transcribed speech
additional tagsets: linking, corpus and names and dates.
 -->
<!DOCTYPE TEI.2 SYSTEM "TIC.dtd">
<TEI.2>

 <teiHeader>
   <fileDesc>
     <titleStmt><title>  The  Verona  Initiative  -  an  electronic
          transcription</title>
     <respStmt><resp>transcribed  and  encoded:</resp><name>Marco
          Cencini</name></respStmt></titleStmt>
     <extent>words: 1869 kb: 14</extent>
     <publicationStmt><authority>release  authority:
          SSLMIT</authority><availability  status="free"><p>Available  for
          purposes  of  academic  research  and  teaching
          only</p></availability></publicationStmt>
     <sourceDesc><recordingStmt><recording> <equipment><p>
          Audio recorded during the conference "The Verona
          Initiative".</p></equipment></recording> </recordingStmt>
     </sourceDesc>
   </fileDesc><encodingDesc>
      <classDecl><taxonomy>
       <category id="mod1">
       <catDesc>consecutive</catDesc></category>
       <category id="mod2">
       <catDesc>simultaneous</catDesc></category>
       <category id="mod3">
       <catDesc>chuchotage</catDesc></category>
       <category id="pos1">
       <catDesc>on-screen interpreter</catDesc>
       </category>
```

```
        <category id="pos2">
        <catDesc>off-screen interpreter</catDesc>
        </category>
        </taxonomy></classDecl></encodingDesc>
      <profileDesc><creation><date>
        24 Sept 2000</date>
        </creation><langUsage><language id="eng">
            english</language>
          <language id="ita">
            italian</language></langUsage>
       <particDesc>
        <person id="bCR" sex="f" role="speaker">
            <persName> Cathy Reed</persName>
            <firstLang> English</firstLang></person>
        <person id="iPS001" sex="f"
         role="interpreter">
          <persName> Cristina Mazza </persName>
          <firstLang>Italian</firstLang></person>
       </particDesc>
      <settingDesc><setting>
        <p> The conference was held in Verona
            on 7-8 July 2000.</p>
         <p> This transcription refers to the first speech
            of the open session of the concluding day
            of the conference.</p>
        <p> The speaker was reading from her notes.
            </p>
         <p> Interpreters were not given the texts. </p>
         <p> They were working in groups of two</p>
         <p>The interpretation was delivered through the
            simultaneous mode.</p>
       </setting></settingDesc>
        <textClass><catRef target="mod2"/>
        </textClass></profileDesc>
  </teiHeader>
  <text>
    <body>
      <div1>
  <gap desc="material missing"/>
<u who="bCR" id="u1" lang="eng" corresp="u2">
  thank you <anchor id="s1" synch="s3"/><pause/>
  my name is Cathy Reed and I'm a director of the North
  West England regional development Agency <pause/>
  with ehm lead responsibility for health and regeneration <pause/>
  I chair the North West Health Partnership <pause/>
  and my background is political leader in local
  government <pause/>
```

L'esempio sopra riportato mostra che una via per poter codificare e annotare le trascrizioni in un formato leggibile dal computer si basa sull'inserimento di etichette (*tags*) compatibili con il linguaggio XML – *eXtensible Markup Language*. Uno dei principali vantaggi è dato dalla possibilità di non visualizzare le annotazioni per esteso, pur essendo presenti, ottenendo conseguentemente un testo leggibile anche dall'analista secondo metodi tradizionali. Per contro, un ostacolo deriverebbe dal livello di alfabetizzazione informatica necessario per poter maneggiare agevolmente questo tipo di codifica

(Sinclair 1995). Questa osservazione potrà sembrare totalmente fuori luogo ad alcuni, specialmente se si considera la diffusione di programmi con cui annotare qualsiasi testo con una certa facilità. Ciononostante, non sempre si riesce a possedere completa padronanza di così tante discipline diverse tra loro quante ne sono emerse nel corso della presente trattazione. È questo, forse, un altro chiaro sintomo della necessità di collaborazione interdisciplinare (Gile 1994) come ingrediente di base in un progetto attinente ai CIS.

Un'alternativa all'uso di attributi in formato XML "puro" risiede nell'applicazione di diversi "strati" di annotazione seguendo uno schema modulare, simile a quanto si trova in un database relazionale. Questo è il modello utilizzato nel sistema della CWB – *Corpus Work Bench* (Christ 1994). In realtà, il sistema CWB è in grado di interfacciarsi anche con testi già annotati su base XML, testi appartenenti ad altre fonti organizzate a livello strutturale o indicizzate, riportando automaticamente il tutto alla «hierarchical, modularized system architecture» (*ibid.*, p. 2) che gli è propria. Dei tre strati di cui si compone, quello che più ci interessa descrivere è il *physical layer*, ossia quello che «encapsulates knowledge about file and tool access and provides an interface which is independent of the storage device and the information type» (*ibid.*, p. 3). Le informazioni compatibili con questo strato sono le seguenti: attributi posizionali, attributi strutturali, bigrammi, informazioni sull'allineamento e attributi dinamici. Perché le funzionalità di ricerca possano espletarsi correttamente, i dati (nel nostro caso le trascrizioni con le rispettive annotazioni) devono essere stati previamente elaborati secondo certi requisiti: «character set normalization, tokenization, sentence boundary detection (if required), and – in case of annotated corpora – the partitioning of the different positional attributes […] into several files. Then, a special one-word-per-line format is produced which is used as input for the construction of the internal corpus representation and the indices» (*ibid.*, p. 4). Per fare un esempio concreto, la Figura 1.9 e la Figura 1.10 mostrano, rispettivamente, lo stesso testo annotato, sia secondo il sistema appena presentato, sia secondo un sistema di rappresentazione con *tags* XML (le parole evidenziate in grigio sono i *token* incolonnati, seguiti dal lemma e dalla parte del discorso, annotazioni di cui parleremo in seguito (§1.3.4.1):

Figura 1.9 Esempio di annotazione su base modulare.

allora	allora	ADV
ce	ce	PRO:pers
l'	il	DET:def
ab-	UNKNOWN	TRUNC
ce	ce	PRO:pers
l'	il	DET:def
abbiamo	avere	VER:pres
fatta	fare	VER:pper

Figura 1.10 Esempio di annotazione su base XML.

```
<w tok="allora" lem="allora" pos="ADV" id="1-1-1"> allora </w>
<w tok="ce" lem="ce" pos="PRO:pers" id="1-1-2"> ce </w>
<w tok="l'" lem="il" pos="DET:def" id="1-1-3"> l' </w>
<w tok="ab-" lem="UNKNOWN" pos="TRUNC" id="1-1-4"> ab- </w>
<w tok="ce" lem="ce" pos="PRO:pers" id="1-1-5"> ce </w>
<w tok="l'" lem="il" pos="DET:def" id="1-1-6"> l' </w>
<w tok="abbiamo" lem="avere" pos="VER:pres" id="1-1-7"> abbiamo </w>
<w tok="fatta" lem="fare" pos="VER:pper" id="1-1-8"> fatta </w>
```

A partire dal file annotato secondo uno dei profili compatibili prima citati, l'impostazione modulare che si ha nel sistema della CWB crea poi più file separati, ma "stratificati" e collegati l'uno con l'altro, per ogni livello di annotazione, per cui il sistema di ricerca è in grado di collegare ogni risultato tra i diversi gruppi di file.[44]

Spostando l'attenzione sugli standard di annotazione, invece, Leech (1997a, pp. 6-8) presenta una serie di considerazioni (una sorta di linee guida o buone prassi) che, se seguite, dovrebbero auspicabilmente ridurre i problemi di compatibilità dei dati con i metodi di ricerca utilizzati in altri progetti o da altri ricercatori. Tali indicazioni possono essere così sintetizzate:

1. Deve essere sempre possibile tornare alla versione "grezza" del corpus, cioè al corpus privo di annotazioni, senza particolari difficoltà.
2. Le annotazioni si devono poter estrarre dal corpus, così come togliere e salvare separatamente.
3. Si deve avere accesso alla documentazione inerente al corpus, tra cui lo schema notazionale (con l'elenco e la spiegazione di tutte le annotazioni), informazioni sulla metodologia utilizzata e le persone coinvolte nell'annotazione, nonché il grado di correttezza delle annotazioni stesse.
4. Le annotazioni devono essere viste per quello che sono, ovvero uno strumento utile anche ad altri soggetti potenzialmente interessati all'analisi avanzata di un corpus.
5. I tipi di annotazione applicati devono basarsi su una riflessione teorica generale e non essere solamente determinati dal tipo di analisi che si intende effettuare.
6. Nessuno schema notazionale può dirsi esaustivo o proporsi come standard assoluto.

[44] Una simile impostazione era stata usata già nel primo progetto di corpus orale: «The original SEC contained several parallel files containing the same text annotated with different kinds of information» (Knowles 1993, p. 108).

Le ultime due indicazioni sembrerebbero contraddirsi, poiché alla dimensione "neutra" e generale delle annotazioni, frutto di una riflessione teorica, si contrappone la dimensione specifica, determinata dagli obiettivi propri di una ricerca e dalla dimensione del corpus stesso. Ad ogni modo, parrebbe maggiormente opportuno prediligere la dimensione generale, oltre che per i motivi già espressi (*re-usability* e *mutual exchange*), anche per un semplice fatto di *inertia*: «if you are familiar with some annotation scheme that you have found useful […], it makes sense to stick to that one in developing your own annotated corpus» (Leech 1997a, p. 7).

Come già ribadito più volte, sono tanti i livelli di annotazione possibili, quali l'annotazione grammaticale (Leech 1997b), l'annotazione sintattica (Leech & Eyes 1997), l'annotazione semantica (Wilson & Thomas 1997) e la cosiddetta *discourse annotation* (Garside et al. 1997a).[45] Ulteriori livelli di annotazione comprendono, per esempio, l'annotazione prosodica, l'annotazione pragmatica e l'annotazione stilistica (Leech et al. 1997), così come la lemmatizzazione, cioè l'annotazione del lemma di ciascun *token* (come nella Figura 1.9 prima mostrata), e l'annotazione di tipi di errori commessi da soggetti che scrivono o parlano nella L2, come potrebbe essere fatto nel caso dei *learner corpora* (Leech 1997a, p. 15).[46] Similmente, se si definissero in maniera precisa alcune delle strategie utilizzate dai traduttori e dagli interpreti nel produrre il TA, si potrebbe proporre di applicare un livello speciale di annotazione per segnalare la presenza di particolari strategie traduttive. Un tale livello di annotazione potrebbe essere chiamato *TS-tagging* o *Translation Strategy annotation*.

Tra i vari tipi di annotazione menzionati, approfondiremo quelli che sono stati indicati come livello minimo per i corpora di interpretazione (Armstrong 1997) e che, in generale, si ritiene possano dare un considerevole valore aggiunto a un corpus per lo studio della lingua.

[45] Per un esempio di corpus scritto con quest'ultimo tipo di annotazione si veda lo studio di Carlson et al. (2001).

[46] Un sistema di annotazione che si avvicina a questo è stato proposto con il nome di approccio MRC – *Meaning, Rhetorical value, Clarity* (Lindquist 2005, Lindquist & Miguélez 2006, pp. 105-106) ai fini dell'autovalutazione delle *performance* in simultanea da studenti interpreti. Il sistema prevede l'applicazione di codici alfanumerici a unità testuali con cui indicare la posizione e il tipo di "errore" commesso o "meccanismo" messo in atto per gestire difficoltà traduttive. Nonostante l'utilità didattica dimostrata, questo sistema appare piuttosto complicato e sembrerebbe richiedere uno sforzo smisurato in termini di tempo (non è chiaro se l'annotazione sia effettuata in modo automatico o interamente manuale).

1.3.4.1 L'annotazione grammaticale e la lemmatizzazione

Negli esempi di codifica illustrati nella Figura 1.9 e nella Figura 1.10 sono inclusi alcuni livelli specifici di annotazione, quali la lemmatizzazione e la parte del discorso, detta anche *POS-tagging* (*Part of Speech*). Come è stato accennato all'inizio di questo capitolo, l'annotazione di queste informazioni attraverso l'applicazione di etichette specifiche può essere effettuata automaticamente, con l'impiego di appositi programmi chiamati *taggers*. Essi sono in realtà indispensabili, in quanto l'inserimento manuale di ogni singola etichetta grammaticale sarebbe un'operazione impraticabile anche in un corpus di dimensioni contenute. Questo non significa che i *taggers* non commettano errori, specialmente nel caso in cui il testo da annotare sia una trascrizione (Moreno Sandoval & Guirao 2003; Marcos Marín 2005, pp. 89-92). I programmi di annotazione automatica si basano, infatti, su una serie di regole interne (un dizionario e una grammatica) o su algoritmi probabilistici per "decidere" quale etichetta assegnare a ogni singola parola (Jurafsky & Martin 2000, pp. 287 e ss.). È evidente che qualsiasi insieme di regole presenterà sempre dei limiti, come nel caso in cui alcune parole non siano state incluse; allo stesso modo, l'approccio stocastico o probabilistico non sempre riuscirà a gestire l'andamento "irregolare" della lingua parlata (si pensi alle ripetizioni, alle false partenze e così via). Rimane la possibilità di correggere manualmente le occorrenze etichettate in modo erroneo dal *tagger* e, successivamente, di impiegare le trascrizioni annotate correttamente come un *training corpus*, vale a dire uno strumento con cui "allenare" il *tagger* e migliorarne la resa (Leech 1997a, p. 9).

Affinché il risultato finale sia accettabile, è chiaro che esso dovrebbe essere il frutto della combinazione di un'annotazione automatica e un'annotazione manuale, in cui editare e correggere eventuali imperfezioni. Si tenga presente che la correzione manuale può essere agevolata dall'estrazione automatizzata dei *token* più problematici (McEnery & Rayson 1997).

Come è stato notato precedentemente, nell'ambito dei CIS, tra le infinite possibilità di annotazione, sempre Armstrong (1997) ribadisce quanto già espresso in merito alla trascrizione, cioè la necessità di attenersi, almeno in un primo momento, a un livello essenziale. A questo proposito, sottolinea proprio l'utilità del *POS-tagging*. Ovviamente, esistono ulteriori tipi e livelli di annotazione, anche più avanzati, ma come punto di partenza il *POS-tagging* sembrerebbe essere sufficiente per poter esplorare il corpus in un modo altamente proficuo.[47]

[47] Per i corpora orali, Leech (1997b) suggerisce anche l'annotazione dei segnali discorsivi e delle esitazioni. Nel primo caso, si renderebbe probabilmente necessario formare le cosiddette *multiwords*, cioè

Esistono diversi programmi per realizzare questa operazione, ma va precisato che strumenti simili non esistono nella stessa misura per tutte le lingue, come dimostrato dalla forte disparità che emerge se si confronta la situazione dell'inglese rispetto all'ebraico (Shlesinger & Ordan 2010). Uno degli etichettatori grammaticali-morfologici più conosciuti è chiamato *Treetagger* (Schmid 1994, 1995), attualmente disponibile in più versioni per diverse lingue: inglese, tedesco, francese, italiano, spagnolo, bulgaro e russo. Per ogni lingua esiste un *tagset* specifico, cioè un repertorio di etichette che esprimono un certo numero di informazioni (per esempio, funzione grammaticale, genere, numero e così via). Nella Tabella 1.10 è riportato il tagset per la lingua inglese:

Tabella 1.10 Tagset di Treetagger per la lingua inglese.

CC	Coordinating conjunction
CD	Cardinal number
DT	determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past principle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

"unire" i token che compongono un unico segnale discorsivo, come è stato riportato per il corpus spagnolo del progetto C-ORAL-ROM (§1.2.1).

Per quanto riguarda la lingua italiana, sono disponibili due versioni. La seconda versione (Tabella 1.12) è stata messa a punto da Baroni et al. (2004) ed è un'estensione della prima (Tabella 1.11):

Tabella 1.11 Tagset di Treetagger per la lingua italiana (versione standard).

ABR	abbreviation
ADJ	adjective
ADV	adverb
CON	conjunction
DET:def	definite article
DET:indef	indefinite article
INT	interjection
NOM	noun
NPR	name
NUM	numeral
PON	punctuation
PRE	preposition
PRE:det	preposition+article
PRO	pronoun
PRO:demo	demonstrative pronoun
PRO:indef	indefinite pronoun
PRO:inter	interrogative pronoun
PRO:pers	personal pronoun
PRO:poss	possessive pronoun
PRO:refl	reflexive pronoun
PRO:rela	relative pronoun
SENT	sentence marker
SYM	symbol
VER:cimp	verb conjunctive imperfect
VER:cond	verb conditional
VER:cpre	verb conjunctive present
VER:futu	verb future tense
VER:geru	verb gerund
VER:impe	verb imperative
VER:impf	verb imperfect
VER:infi	verb infinitive
VER:pper	verb participle perfect
VER:ppre	verb participle present
VER:pres	verb present
VER:refl:infi	verb reflexive infinitive
VER:remo	verb simple past

Tabella 1.12 Tagset di Treetagger per la lingua italiana (versione ampliata).

ADJ	adjective
ADV	adverb (excluding -mente forms)
ADV:mente	adveb ending in -mente
ART	article
ARTPRE	preposition + article
AUX:fin	finite form of auxiliary
AUX:fin:cli	finite form of auxiliary with clitic
AUX:geru	gerundive form of auxiliary
AUX:geru:cli	gerundive form of auxiliary with clitic
AUX:infi	infinitival form of auxiliary
AUX:infi:cli	infinitival form of auxiliary with clitic
AUX:ppast	past participle of auxiliary
AUX:ppre	present participle of auxiliary
CHE	che
CLI	clitic
CON	conjunction
DET:demo	demonstrative determiner
DET:indef	indefinite determiner
DET:num	numeral determiner
DET:poss	possessive determiner
DET:wh	wh determiner
NEG	negation
NOCAT	non-linguistic element
NOUN	noun
NPR	proper noun
NUM	number
PRE	preposition
PRO:demo	demonstrative pronoun
PRO:indef	indefinite pronoun
PRO:num	numeral pronoun
PRO:pers	personal pronoun
PRO:poss	possessive pronoun
PUN	non-sentence-final punctuation mark
SENT	sentence-final punctuation mark
VER2:fin	finite form of modal/causal verb
VER2:fin:cli	finite form of modal/causal verb with clitic
VER2:geru	gerundive form of modal/causal verb
VER2:geru:cli	gerundive form of modal/causal verb with clitic
VER2:infi	infinitival form of modal/causal verb
VER2:infi:cli	infinitival form of modal/causal verb with clitic
VER2:ppast	past participle of modal/causal verb
VER2:ppre	present participle of modal/causal verb
VER:fin	finite form of verb
VER:fin:cli	finite form of verb with clitic
VER:geru	gerundive form of verb
VER:geru:cli	gerundive form of verb with clitic
VER:infi	infinitival form of verb
VER:infi:cli	infinitival form of verb with clitic
VER:ppast	past participle of verb
VER:ppast:cli	past participle of verb with clitic
VER:ppre	present participle of verb
WH	wh word

Assieme all'annotazione grammaticale, anche la lemmatizzazione darebbe un notevole valore aggiunto al corpus. Essa consiste nell'abbinare un'etichetta

indicante la parola nella sua forma base a ciascun *token* (si vedano gli esempi riprodotti alla Figura 1.9 e alla Figura 1.10). I procedimenti di etichettatura automatica funzionano per tutte le parole incluse nella grammatica interna di ogni *tagger*. Nel caso di parole non contemplate in tale grammatica, il *tagger* potrebbe applicare una etichetta "neutra" (per esempio, UNKNOWN) per segnalare la situazione; diversamente, potrebbe essere applicata l'etichetta che si avvicina alla soluzione probabilisticamente più accettabile.

Una delle principali ragioni per cui vale la pena etichettare in questo modo un corpus è che attraverso l'uso dei *POS-tags* e della lemmatizzazione è possibile ampliare le potenzialità di ricerca semiautomatica. Per fare un esempio, si consideri il caso di una lingua morfologicamente ricca come l'italiano, dove, tra le altre cose, esistono diverse forme verbali a seconda della persona.[48] Ad esempio, se volessimo verificare la frequenza e l'uso del verbo "cantare" in un corpus italiano senza lemmatizzazione e *POS-tagging*, dovremmo svolgere tante ricerche quante sono le possibilità di coniugare tale verbo, nelle diverse persone, così come nei diversi tempi e modi. In realtà, grazie alla presenza dell'etichetta riportante il lemma o la parte del discorso, sarebbe sufficiente svolgere una sola ricerca di tale *tag* per ottenere come risultato l'elenco di tutte le occorrenze del verbo in questione, indipendentemente dalla forma (declinazione) particolare in cui possa apparire nel corpus.

Nella Linguistica computazionale non vi è sempre corrispondenza tra ciò che è considerato una "parola" e ciò che in termini tecnici è chiamato "*token*". I *token* sono le singole parole di cui si compone un corpus; quando un corpus è "tokenizzato" significa che ogni singola parola occupa una riga, per cui tutte le parole sono mandate a capo e risultano incolonnate una sotto l'altra. Tuttavia, ciò che si fa rientrare nel concetto di "parola" potrebbe comprendere più *token*, come nel caso dei nomi composti. Un esempio tipico è dato dal nome proprio "New York": a ben vedere esso è composto da due elementi; a seconda dell'approccio adottato, saranno considerati due *token* (che costituiscono una parola) o due parole. Un altro esempio è dato dalle parole apostrofate, come "l'altro" in italiano o "I'm" in inglese. Se si utilizza la funzione di calcolo del numero di parole presente nei programmi di videoscrittura (nel nostro caso abbiamo testato sia MS Word, sia TextPad), i due casi proposti risultano essere costituiti da una sola parola. Tuttavia, operando la tokenizzazione i due casi apparirebbero "spezzati" in due elementi ciascuno, ovvero due *token*: l' + altro (per l'esempio in italiano) e I + 'm (per l'esempio in inglese). Questa puntualizzazione è di fondamentale importanza nel presentare i dati contenuti in

[48] Questo spiega il motivo per cui il numero di etichette comprese in un *tagset* varia anche a seconda della lingua. Per esempio, è stato stimato che per l'inglese sono stati compilati *tagset* che vanno da 30 a 200 etichette, mentre per lo spagnolo si arriva fino a 475 (Leech 1997b).

un corpus, poiché il numero di parole potrebbe non coincidere con il numero di *token*.

Un'ulteriore distinzione fondamentale nell'ambito della Linguistica computazionale è tra i concetti di *token* e *type*. Se il numero di *token* è ottenuto dal calcolo delle loro occorrenze totali all'interno di un corpus, con i *type* si calcola il numero di occorrenze dei *token* che appaiono con la stessa forma grafica (Jurafsky & Martin 2000, p. 195). Ad esempio, nella frase "mi piace questo e mi piace quello" si avranno in totale sette *token* e cinque *type* ("mi" e "piace" si ripetono due volte). Il calcolo del rapporto tra il numero di *type* e di *token* è una misurazione "classica" in linguistica computazionale (*type-token ratio*), poiché indice del grado di varietà lessicale in un corpus (Stubbs 1996, Castello 2004).

In linea di massima, si riscontra una corrispondenza diretta tra la rappresentazione ortografica e la dimensione morfosintattica di ogni *token*. Tuttavia, esistono anche delle eccezioni, come nel caso delle *multiwords*, dei clitici (per esempio in "venderlo", "chiamatelo") e nelle parole composte come "*world-class*", "*twenty-seven*", ecc. (Leech 1997b). A seconda del *tagger* utilizzato, questi casi possono essere trattati come *token* individuali, oppure possono essere "decomposti" di modo che i singoli componenti siano etichettati individualmente.

Come si nota dalle liste di etichette (*tagset*) presentate nelle Tabelle 1.10, 1.11 e 1.12, il nome impiegato per un'etichetta grammaticale è spesso il risultato di un'abbreviazione piuttosto intuitiva. A questo proposito, Leech (1997b) ha indicato alcuni criteri da seguire nella scelta del nome da assegnare ai *tag* (non solo grammaticali): *conciseness* (brevi), *perspicuity* (devono cioè essere facili da ricordare), *analysability* (devono consentire l'analisi automatica) e *disambiguity* (non devono essere ambigui).

1.3.4.2 Segmentazione e unità di analisi

Un altro passo fondamentale da effettuare in fase di trascrizione, annotazione e codifica di un corpus orale al fine di poterne studiare i dati è la suddivisione del flusso comunicativo in segmenti o unità di analisi. Stando alle considerazioni teoriche sopra esposte, l'uso della punteggiatura risulterebbe fuorviante e, in ogni modo, sarebbe condizionato eccessivamente dalla sola percezione soggettiva del trascrittore (Blanche-Benveniste 2005, pp. 51-52): come distinguere l'annotazione di unità attraverso l'uso del punto, del punto e virgola,

dei due punti, e così via? Come si potrebbe quantificare il valore di ogni segno di interpunzione per poi esprimerlo in unità discrete?[49]

L'alternativa "orale" alla punteggiatura è l'annotazione prosodica, in cui «prosodic symbols are representations of part of a spoken transcription, indicating the way in which a piece of spoken language was uttered» (Leech et al. 1997, p. 85). Tuttavia, questo livello di annotazione sembrerebbe richiederebbe uno sforzo considerevole, pertanto sarebbe con buona probabilità più semplice da applicare in seconda battuta, utilizzando cioè un corpus già disponibile, le cui trascrizioni presentano "solo" un livello minimo di annotazione. Tale livello minimo potrebbe limitarsi a una prima segmentazione del testo trascritto (senza però scendere nei dettagli di un'annotazione prosodica, in cui sarebbe doveroso rendere conto dell'intera gamma di variazioni dell'andamento prosodico attraverso un'analisi spettrografica). Partendo dall'estremo opposto, si tratterebbe insomma di «individuare le unità di analisi del parlato superiori a quelle fonetiche e morfologiche e correlative alle aggregazioni di "contenuto"; unità, evidentemente, inferiori a quelle di testo […]» per così «[…] riscontrare l'esistenza di tipi di organizzazione sintattico-semantica propri del parlato» (Nencioni 1989, p. 244). Una tale segmentazione non solo è utile ai fini dell'analisi, ma renderebbe anche la trascrizione più accessibile all'analista che volesse leggerla in modo tradizionale, stampandola su carta o visualizzandola direttamente su uno schermo.

> Segmentation is the separation of a section of discourse based on the identification of single utterance acts. For this purpose the person transcribing reconstructs the communicative process from a perspective outside the original speech situation. He or she applies knowledge about communication to the data to find out how the linguistic surface is structured […].
> (Meyer 1998, p. 72)

Come si può constatare anche dalla citazione di Meyer, due dei termini generalmente impiegati per riferirsi alle unità in cui si articola il parlato sono "enunciato" (*utterance*) e "atto" (*act*). Nonostante Meyer li usi assieme per esprimere un unico concetto, di norma il primo è considerato espressione concreta dell'altro, o meglio della "frase", questi ultimi intesi come una entità astratte e teoriche. Diversi studiosi prediligono l'una o l'altra dimensione, mettendo così in luce una certa indeterminatezza della cornice teorica che

[49] L'inclusione della punteggiatura è comunque ammesso in base all'uso che si desidera fare delle trascrizioni. Per esempio, una parte dei materiali spagnoli del corpus C-ORAL-ROM (trascritti senza segni di interpunzione) è stata rivista con l'aggiunta della punteggiatura, al fine di creare materiali per la didattica dello spagnolo come L2 (Campillos et al. 2007). Si veda anche la proposta di Setton (2002) in merito alla *fluent version* di una trascrizione (§1.3.3.2).

stiamo discutendo. Per esempio, tra le possibili unità linguistiche del parlato, Cresti (2000a, cap. 4) parte dall'individuazione della "battuta". Tuttavia, precisa subito che la battuta non può essere considerata un'unità operativamente utile, poiché va da una sola parola o interiezione a una lunga argomentazione. Pertanto, orienta la sua indagine mettendo a fuoco gli atti comunicativi,[50] intesi come le attività attraverso cui si esplica il parlare, facendo leva sul carattere "affettivo" del parlato, legato cioè all'agire (atti) attraverso unità linguistiche (enunciati). In questo senso, l'enunciato corrisponderebbe alla realizzazione linguistica di un atto. Sul piano teorico, per delineare i confini di un enunciato si ritiene che debbano essere soddisfatti complessivamente i seguenti requisiti: intonazione, predicazione, compiutezza semantica e autonomia (*ibid*., p. 57). In particolare, l'intonazione giocherebbe un ruolo chiave al consentire di determinare le unità tonali all'interno degli enunciati stessi, conformemente al principio secondo cui «Il giudizio sulla interpretabilità pragmatica di un evento orale è basato sulla valutazione dei suoi indici intonativi» (Moneglia 2005, p. 218). Le unità tonali si distinguono a loro volta in unità tonali terminali e non terminali, annotabili con l'inserimento di una doppia barra (//) o una barra singola (/) rispettivamente (*ibid*., pp. 218-219; Cresti 2000a, pp. 38-39). Si noti che è stata menzionata l'interpretabilità pragmatica, e non semplicemente semantica. Di fatti, agli enunciati è attribuito un carattere "autonomo" fintanto che con autonomia si intende l'interpretabilità pragmatica (distinta dalla compiutezza semantica, che compete alla frase). A seguito delle analisi sui corpora orali da lei esaminati, secondo Cresti l'enunciato in realtà «è identificato dai requisiti di intonazione e autonomia e non da quelli di predicazione e compiutezza semantica, che sono propri invece della frase [...]»; inoltre, «come espressione semantica piena, semplice [...] o complessa [...], intonata secondo un comment, è autonoma[sic] perché consente la propria interpretabilità pragmatica» (*ibid*., p. 59).[51] In definitiva,

> […] un enunciato può essere un pattern informativo complesso, costituito da un'unità informativa di Comment più altre unità informative di funzione diversa, e in tal caso è interpretato da un pattern tonale complesso, costituito da più unità tonali intorno ad un'unità tonale di Comment.
> (Cresti 2000a, p. 118)

Tra le varie unità abbinabili al Comment, sono menzionati i seguenti tipi: topic, appendici, ausilii dialogici, incipit, fatici, allocutivi, conativi, incisi, introduttori locutivi, commenti multipli. Senza ora scendere nel dettaglio di ogni singola

[50] Diversi dagli atti linguistici, utilizzati invece nello studio di Bilbow (2007).
[51] "Comment" è definito come «quell'espressione che serve a compiere l'illocuzione» (*ibid*., p. 81).

unità, il risultante quadro generale vedrebbe l'organizzazione dei testi orali articolarsi in *enunciati* e *unità di informazione*, secondo una «sintassi segmentata» (*ibid*., p. 168 attingendo da Bally 1971) con uno schema Tema-Rema – in opposizione alla sintassi legata della lingua scritta che seguirebbe strutture di tipo logico.[52] Con questo, non si deve pensare che un simile approccio alla strutturazione del parlato non possa essere applicato anche ai formati interazionali dialogici. Al contrario, la dimensione di "enunciato", quale che sia la sua natura, semplice o complessa, è sempre individuabile all'interno di singole battute (in altre parole, si va a destrutturare un'interazione dialogica nel suo complesso in tanti micro interventi individuali).[53] A questo proposito, è interessante notare che nel campione studiato da Cresti «nei testi di piena spontaneità la maggioranza degli enunciati è semplice e in quelli con forme di programmazione invece è complessa» (2000a, p. 175). La natura complessa degli enunciati trova espressione nelle particolari relazioni sintattiche che si instaurano con altre unità (e non all'interno della stessa unità). Tali relazioni rientrano in un'organizzazione complessiva che è stata definita come macrosintassi e interessano unità di informazione diverse, riferibili appunto agli enunciati complessi.

A questo punto è doverosa una precisazione. L'uso di termini come "complesso" o "segmentato" (così come altri che saranno ripresi a breve) al fine di proporre una concettualizzazione dei modi in cui si articola il parlato e, conseguentemente, elaborare strumenti funzionali all'analisi (come lo possono essere appunto gli enunciati) rischia di spingere univocamente verso la tradizionale concezione secondo cui "comunicazione parlata" è sinonimo di caos, disordine e mancato rispetto delle regole, in opposizione a quanto si avrebbe nella comunicazione scritta. A differenza di quanto avveniva in passato, attualmente non manca di certo l'interesse accademico nei confronti della comunicazione parlata, ma vale la pena cercare di dare il giusto risalto ai lavori di chi è riuscito a rendere conto di modalità in cui le informazioni sono

[52] La trattazione di questo argomento da parte degli stessi autori Cresti e Moneglia è disponibile anche in lingua inglese (Cresti 1995, Moneglia & Cresti 2001).

[53] In ambito ispanista è stata formulata una proposta simile, ma con una terminologia alternativa (Hidalgo & Padilla 2006). Le unità individuate sono chiamate *actos* e *subactos*. Le prime «son unidades jerárquicamente inferiores a las intervenciones que poseen las propiedades de la *identificabilidad* y de la *aislabilidad*» (*ibid*., p. 118), mentre le seconde sono «subunidades integrantes del Acto, reconocibles como aportes o soportes informativos relevantes, pero no aislables en el contexto dado» (*ibid*., p. 123). L'identificazione e l'isolamento di tali unità si baserebbe, dunque, non solo sulle indicazioni intonative, ma anche sulle informazioni semantico-contenutistiche e contestuali. Anche questo sistema si articola in ulteriori tipi di "unità" così individuate, consentendo apparentemente di gestire anche modalità di espressione che si discostano notevolmente da uno sviluppo "ordinato" e "idealizzato" dell'ordine informazionale. Pur trovando una forte consonanza con le riflessioni illustrate da Cresti (2000a), uno svantaggio della tassonomia proposta da Hidalgo & Padilla (2006) potrebbe forse provenire dal fatto che i termini operativi utilizzati rievocano direttamente la teoria degli atti linguistici (Austin 1976), con il rischio di creare confusione all'interno delle discipline che si rifanno a tali costrutti.

"assemblate" per essere espresse oralmente. Nencioni (1989, p. 244 attingendo da Sornicola 1981) avverte infatti che:

> il testo del parlato presenta macro- e microstrutture diverse da quelle dei testi elaborati, ha cioè una organizzazione non analitica ma "globalizzante", la cui macrostruttura è spesso slogata ed ambigua e ricavabile come un *puzzle*. Esso soffre spesso del collasso delle microstrutture, presenta blocchi informativi olofrastici, ordinamento seriale, uso minimale del linguaggio […].

Ammesso che l'ordine secondo cui sono organizzate le informazioni è differente nel parlato (un *puzzle*) rispetto allo scritto, non suonerebbe forse più interessante dire che desideriamo "affrontare" tale *puzzle* come frutto di un'attività intelligente che racchiude rivelazioni latenti, e non come un problema involuto in cui regna il disordine? È questo il senso in cui ci sentiamo di sposare l'approccio proattivo presente nella Traduttologia.[54]

L'individuazione delle caratteristiche peculiari da riferire alla strutturazione della comunicazione parlata parte spesso da un'operazione di confronto tra modelli astratti, ovvero idealizzazioni dei testi scritti e dei testi parlati. Abbinando la riflessione teorica all'analisi di dati empirici (per quanto limitati), Sornicola (1981, 1984) ha tentato di fornire una risposta «all'esigenza di reperire le caratteristiche invarianti dei testi parlati, valide cioè per i testi monologici come per quelli dialogici, per quelli narrativi come per quelli argomentativi, per parlanti a competenza sociolinguisticamente alta come per quelli a competenza sociolinguisticamente bassa» (Sornicola 1984, p. 343). Al termine della sua trattazione conclude:

> […] ciò che appare caratteristico dei testi parlati è una *struttura informativa* diversa da quella dei testi scritti. La distribuzione dell'informazione rispetto ai blocchi informativi, sia a livello micro- che a livello macro- è diversa dai testi scritti. Assumendo il testo come uno spazio-tempo, possiamo dire che la tipologia dei testi parlati presenta una costituzione particolare: si determinano, infatti, delle improvvise aree di decrescenza dell'informazione […] o addirittura dei *gaps* informativi repentini […].
> (Sornicola 1984, p. 349)

Quel che è peggio, apparentemente, è che di fronte a cotanta imprevedibilità distribuzionale delle informazioni, né le indicazioni prosodiche, né la presenza delle pause fungono sempre da elementi regolatori, in base ai quali ricostruire la struttura del flusso comunicativo conformemente ai "blocchi" semantici e

[54] Si veda il numero 50/4 della rivista *Meta* (2005).

concettuali espressi (Sornicola 1981, pp. 14-17). A ben vedere, questo è ancora più vero se si considerano l'andamento prosodico e la distribuzione delle pause nei TA prodotti dagli interpreti in modalità simultanea. Le particolari condizioni di produzione del TA favoriscono, infatti, la presenza di fenomeni innaturali, e comunque diversi da quanto avverrebbe se la produzione del testo non dipendesse da un input esterno e prodotto da altri (si vedano, tra gli altri, gli studi di Shlesinger 1994, Cecot 2001, Garwood 2002, Ahrens 2005).[55] In aggiunta ai tratti intonativi e paralinguistici peculiari dei TA, nell'accostare testi paralleli tra due o più lingue diverse ci si scontra inevitabilmente anche con diversi gradi di somiglianza morfologica e strutturale di ciascun codice. Per fare un esempio, Vuorikoski (2004) ha dovuto affrontare simili sfide nel tentativo di allineare un TP con tre TA prodotti contemporaneamente in diverse lingue (§2.1). Utilizzando un formato tabulare per la trascrizione ha segmentato i testi in unità «where one line consisted of one unit carrying relevant semantic information», dove «the smallest unit is the propositional utterance (often a clause)» (*ibid.*, p. 104). Tuttavia, l'autrice si è imbattuta in non poche difficoltà a causa delle numerose differenze tra le quattro versioni di uno stesso discorso. Conseguentemente, «in order to obtain a clear view of the degree of correspondence between the STs and the TTs it seemed advisable to focus on the predicate, subject and object. Depending on their informational weight they are either placed on separate lines or separated by a slash [/]» (*ibid.*).

Senza addentrarsi nella varietà specifica del "parlato-interpretato", Sornicola (1981) si è mostrata comunque determinata a «giustificare l'interesse per una definizione di una unità di analisi più ampia della parola, corrispondente, sul piano empirico, al concetto teorico di frase [...]» per capire «come categorie e funzioni del discorso, elaborate secondo sequenze linguistiche idealizzate, si adattassero a sequenze linguistiche reali» (*ibid.*, p. 18). Poiché anche «la nozione di relazione sintagmatica non era generalmente applicabile ai dati» (*ibid.*, p. 19), la sua scelta operativa si traduce nel «condurre l'analisi dei testi in base a sequenze di elementi non necessariamente coincidenti né con la nozione di "frase" né con quella di "enunciato" […], ma di volta in volta determinate secondo il loro interesse ai fini dell'esame del testo» (*ibid.*).

Riportando la nostra riflessione all'ambito della trascrizione dei dati, la conclusione a cui giunge Sornicola risponde efficacemente alle considerazioni pratiche suesposte (§1.3.3.2), tra cui rientrano non solo la necessità di analizzare automaticamente i dati, ovvero fare sì che essi siano leggibili dal computer, ma anche la necessità di poterli leggere "ad occhio nudo" e di annotarli in modo agevole.

[55] Per l'interpretazione in modalità consecutiva standard si vedano gli studi di Mead (2000, 2002a).

1.3.5 Allineamento

La penultima tappa prima di completare il percorso di costruzione di un corpus di interpretazione (parallelo) concerne la questione dell'allineamento. Vi sono due diverse prospettive da cui inquadrare tale argomento, una intratestuale e l'altra intertestuale. La prospettiva intratestuale riguarda la natura orale di un corpus e prevede l'allineamento testo-suono, cioè la possibilità di collegare la rappresentazione dei dati in forma (tra)scritta alla corrispondente registrazione audio/video. Dall'altra parte, la prospettiva intertestuale riguarda la natura parallela di un corpus e prevede l'allineamento tra TP e TA (eventualmente, anche l'allineamento tra più TA riferiti allo stesso TP). Esamineremo entrambe le prospettive, fornendo, come per le altre tappe, considerazioni sia teoriche, sia pratiche.

1.3.5.1 Allineamento testo-suono

Le maggiori questioni teoriche sollevate dalla rappresentazione in forma trascritta del parlato sono state discusse precedentemente (§1.3.3.1) Tra queste, abbiamo messo in luce che la soppressione di numerosi tratti dell'oralità dal sistema notazionale impiegato in una trascrizione potrebbe creare non poche riserve, in quanto l'analista non sarebbe più in grado di "sentire con gli occhi" l'effettiva *performance* fissata in forma scritta, rischiando in questo modo di disorientarsi nell'interpretazione dei dati. Tali riserve si potrebbero ridimensionare in sede di analisi e accesso al corpus, se dal dato trascritto fosse possibile collegarsi direttamente al dato registrato (audio o video). Non che questo sia totalmente esente da problemi di soggettività (le registrazioni audio non permettono di considerare molti tratti del linguaggio non verbale, mentre le registrazioni video offrirebbero una visione della situazione comunicativa limitata a quanto riesce a catturare l'obiettivo della videocamera), ma consentirebbe senza dubbio di avvicinarsi ai dati in maniera molto più completa rispetto al solo testo trascritto, per quanto arricchito di annotazioni dell'oralità.

> Although the possibilities offered by computerized corpora of spoken discourse for advancing understanding are beyond question both exciting and promising, it would be self-defeating to suppose that the problems of the relationship between transcriptions and original speech events are solved by the storage of transcriptions on computer. To take full advantage of the opportunity offered by computerized corpora, we need to intensify, rather than sidestep, our scepticism about this relationship.
> (Cook 1995, p. 35)

Gli stessi ricercatori che si sono occupati del primo corpus orale ad essere realizzato hanno avvertito chiaramente l'importanza di non affievolire tale diffidenza nei confronti della rappresentazione scritta del parlato. Infatti, a conclusione della prima fase di creazione del SEC (*Spoken English Corpus*, Knowles 1993), si erano proposti appunto di allineare il testo al suono: «Instead of linking the text to prosodic transcription, we will need to find a way of linking it via the transcriptions to the waveform itself» (Knowles 1993, pp. 118-119). È evidente che il fatto di poter contare su una trascrizione "alleggerita" delle annotazioni prosodiche comporta vantaggi estremamente interessanti, tra cui uno snellimento nel compito di trascrizione, così come una riduzione delle difficoltà nel trattamento automatico dei testi annotati e codificati. Tuttavia, l'ancoraggio del testo trascritto alla parte di registrazione corrispondente implica che sia effettuato comunque uno speciale tipo di annotazione, attraverso l'inserimento di etichette riportanti un codice temporale per ogni segmento interessato.

Di nuovo si presenta la questione della segmentazione, questa volta non propriamente del testo trascritto, ma della traccia audio. Si noti che questo tipo di segmentazione risponde più che altro a esigenze di supporto all'analisi e non tanto legate alla rappresentazione scritta dei dati. L'allineamento testo-suono può infatti essere eseguito rispetto a ogni singola parola (con molta difficoltà), oppure in corrispondenza degli enunciati o di unità la cui estensione è stabilita dal ricercatore. Ad ogni modo, anche questa scelta dipenderebbe sempre dal principale obiettivo per cui è stato costruito il corpus. Hofland (2003) presenta tre diversi esempi di progetti in cui è stato effettuato l'allineamento testo-suono seguendo differenti metodologie. Nel caso dell'allineamento del suono per ogni singola parola «files were sent to a company in England (SoftSound) for automatic text and sound alignment» (*ibid.*, p. 330). Pertanto, una opzione sarebbe quella di esternalizzare l'intera operazione di allineamento testo-suono, ammesso che le risorse economiche a disposizione del gruppo di ricerca o del singolo ricercatore lo consentano. Oltre a questo limite (non indifferente se si considera lo stato dei finanziamenti alla ricerca, non solo in Italia), si correrebbe anche il rischio di trovare degli errori di allineamento che dovrebbero poi essere corretti manualmente. Nel secondo esempio, le etichette temporali (qui chiamate *time stamp*, ma ci si riferisce ad esse anche come *time code* o *time tag*) sono state tutte inserite manualmente ad intervalli di dieci secondi. Infine, nel terzo esempio, l'annotazione temporale è stata effettuata automaticamente in fase di trascrizione di ogni segmento, attraverso un software apposito chiamato Praat (Boersma & Weenink 2001). «This program keeps track of the time codes for the beginning and end of each segment which is transcribed. The text and the time codes can be read by other programs and are converted to the main format used for searching […]» (Hofland 2003, p. 330).

Dagli esempi riportati si possono ricavare almeno due elementi particolarmente rilevanti alla discussione sull'allineamento testo-suono in un corpus orale. Il primo è che questo tipo di allineamento richiede l'annotazione dei codici o *tag* temporali, i quali devono essere inseriti manualmente all'interno del testo della trascrizione durante o dopo la produzione della trascrizione stessa. Il secondo elemento è che al fine di ancorare il testo alla registrazione (o viceversa) si devono utilizzare programmi informatici appositi. L'esempio citato prima è il software Praat, un programma sviluppato per analisi di tipo fonetico; ciononostante, abbiamo rilevato che attraverso la funzione di annotazione del tempo di inizio e fine dei segmenti trascritti, esso consente non solo di portare a termine l'annotazione temporale ma anche, cosa forse più importante, di esportare i codici temporali in modo da utilizzarli anche con altri programmi. Si ricordi che i principi menzionati precedentemente su come dovrebbe essere applicato un livello di annotazione restano sempre validi, e quindi la compatibilità delle trascrizioni con altri sistemi e programmi dovrebbe essere garantita anche nei confronti dell'annotazione temporale.

Un altro programma che presenta la funzione di inserimento di codici temporali per agganciare un file di testo a una traccia audio o video è Transana. Questo programma è in grado di gestire fino a tre file audio/video contemporaneamente;[56] le etichette temporali sono inserite direttamente nel testo trascritto durante l'ascolto o la visione dei dati multimediali con un comando dato con la tastiera (Ctrl+t). Successivamente si può cliccare direttamente su una porzione di testo perché il riproduttore audio/video si posizioni nel punto esatto corrispondente. Un altro programma simile, ma con funzioni aggiuntive (come il controllo della velocità di riproduzione della traccia audio, quindi non solo stop e play, e l'analisi spettrografica) è WinPitch (Martin 2004). Con questo programma si possono inserire le etichette temporali cliccando con il cursore direttamente nel testo della trascrizione mentre si ascolta la registrazione. In questo modo, la trascrizione risulta segmentata in un formato compatibile per l'esportazione in XML o Excel.

Una modalità di annotazione temporale leggermente diversa si riscontra nei software compatibili con un formato di trascrizione a spartito, o comunque nei software in cui l'annotazione risulta essere effettuata a partire dalla traccia audio/video e non tanto dalla trascrizione. In questi, la traccia multimediale viene ancorata direttamente alla linea dedicata alla trascrizione esattamente nei punti in cui questa è segmentata dall'analista; è altresì possibile stabilire ancoraggi con eventuali altri strati di annotazione che si desidera inserire (da abbinare sempre a una unità di segmentazione). Un esempio è il programma

[56] Questo è stato verificato nella versione demo 2.41 scaricabile da Internet; la stessa funzione non era presente nella versione 2.12 utilizzata nel presente lavoro, la quale consente comunque di associare più file di testo allo stesso file multimediale.

ADAALab (Yagi 1994, 1999), sviluppato ai fini della cosiddetta *Digital Discourse Analysis* dell'interpretazione simultanea. Con ADAALab si possono visualizzare le due tracce sonore del TP e del TA sincronizzate, riuscendo così facilmente a individuare i punti in cui vi sono sovrapposizioni complete, parziali o nulle tra oratore e interprete. Si possono inoltre aggiungere diversi strati o livelli di annotazione, tra cui un livello con la trascrizione vera e propria e altri con qualsiasi tipo di *tag* si desideri considerare. Il programma è poi in grado di estrarre informazioni dai dati inseriti e svolgere calcoli statistici. Uno dei pregi di questo programma è dato dalla precisione con cui è possibile inserire alcuni tipi di annotazione,[57] diversamente da quanto si riuscirebbe a fare con un'annotazione manuale e basata solamente sull'ascolto. Infatti, il programma permette di visualizzare sullo schermo le onde sonore del TP e del TA, supportando così il compito di annotazione con un doppio canale percettivo (la traccia audio assieme alla rappresentazione grafica dell'onda sonora), oltre a quello fornito dalla trascrizione (che andrebbe aggiunta prima di effettuare l'annotazione). Un altro esempio simile è il programma appositamente sviluppato per l'annotazione del corpus multimediale *Forlixt 1* (Valentini 2009). Va precisato che l'annotazione di materiali filmici comporta sostanziali differenze rispetto a materiali tratti da eventi comunicativi mediati da interpreti simultaneisti, a partire dalle unità di segmentazione (evidentemente anche l'obiettivo di una ricerca gioca sempre un ruolo determinante). Nel corpus creato da Valentini si adottano la "scena" e la "battuta" come unità di analisi, segmentando il file multimediale, cioè il film, con codici temporali che creano le sequenze video su cui applicare diversi livelli di annotazione.[58] Il programma (con la sua interfaccia chiamata MovieDB) è uno strumento decisamente innovativo per il campo di studi sulla traduzione multimediale e che potrebbe essere adattato per l'annotazione di materiali tratti da altre situazioni comunicative. Un altro software disponibile per l'allineamento testo-suono è SpeechIndexer (Szakos & Glavitsch 2004a, 2004b), sviluppato in risposta alla necessità di preservare lingue di diffusione limitata e che rischiano di essere dimenticate. Il sistema SpeechIndexer ha due applicazioni principali, una per indicizzare le tracce audio in rapporto ai brani di testo (con una prima

[57] Nel suo studio, effettuato secondo un paradigma cognitivo, Yagi (1999) ha analizzato i *time-management patterns* degli interpreti simultanei in termini qualitativi e quantitativi, in base all'idea secondo cui «the time structure of an interpretation speech signal can reveal a lot of information about its quality» (*ibid.*, p. 273).

[58] Spiega, infatti, Valentini (2009) che il tipo di oralità presente nel genere filmico «preserva […] in misura maggiore le unità informative, veicolate mediante confini prosodici netti (tipicamente da silenzio a silenzio del parlante), in virtù di una struttura sintattica più assimilabile allo scritto. […] La tecnica della scrittura cinematografica, che trae le sue origini da quella teatrale, obbedisce, infatti, a una certa tradizione che esige coerenza, connessione logica tra le battute, e appare subordinata a convenzioni di rappresentazione scenica specifiche, ancorate in primis a esigenze di garanzia di trasmissione dell'informazione» (*ibid.*, p. 79).

segmentazione automatica della traccia audio in base all'andamento dell'onda sonora), e una per svolgere ricerche all'interno della banca dati risultante dal processo di indicizzazione.

Si tratterebbe di valutare fino a che punto i dati annotati con i tre software, tanto ADALaab quanto MovieDB e SpeechIndexer, possono essere esportati su altre piattaforme e risultare compatibili con altri strumenti di interrogazione e di annotazione. È quanto sembra garantire l'ultimo esempio di questa rassegna, ovvero il pacchetto software Exmaralda (Schmidt 2001, 2003, 2004, 2009), nel quale sono compresi un editor per realizzare trascrizioni a sparito (Partitur Editor), un programma di gestione dati (Corpus Manager) e uno strumento di consultazione del corpus (EXAKT – *EXMARaLDA Analysis and Concordancing Tool*). Anche in questo caso, le annotazioni sono riferite alla traccia audio/video attraverso l'ancoraggio a una *timeline* condivisa con la linea dedicata alla trascrizione. Questa risulta quindi segmentata in base al tipo di annotazione che si desidera effettuare, potendo poi estrarre ogni occorrenza assieme al rispettivo frammento multimediale. I materiali trascritti e annotati con questo pacchetto software possono anche essere esportati in un documento html, così come è possibile importare materiali con operazioni di adeguamento a seconda del tipo e del grado di codifica già presenti. Questo programma risulta particolarmente interessante per coloro che si occupano di studiare la comunicazione parlata con un formato interazionale dialogico, ricco di sovrapposizioni e con più partecipanti contemporaneamente. In conclusione, sembra che funzionalità maggiori siano generalmente legate alla specificità di un software costruito *ad hoc*. Tuttavia, dai primi esempi illustrati si evince che è anche possibile provvedere all'annotazione temporale preservando l'interoperabilità dei dati trascritti ed etichettati.

1.3.5.2 Allineamento TP-TA

Oltre all'allineamento testo-suono, tra i requisiti minimi di un corpus di interpretazione avanzati da Armstrong (1997) è menzionato anche l'allineamento tra TP e TA. È questa la prospettiva intertestuale da cui si può affrontare il tema generale dell'allineamento tra TP e TA, una caratteristica di grande valore nei corpora paralleli, poiché «Parallel corpora are most useful when they are aligned, that is, when the texts are matched up so that they can be more easily compared» (Lawson 2001, p. 285). Come è avvenuto anche per le altre tappe nella creazione di un corpus, coloro che si sono occupati dei corpora di lingua scritta hanno per primi affrontato questa sfida (Hofland & Johansson 1998), esplorando diverse strategie atte a produrre un allineamento automatico

di un testo originale con la sua versione tradotta (in una o più lingue, oppure più versioni tradotte nella stessa lingua).[59]

Poiché i sistemi di allineamento automatico si basano sul principio di equivalenza tra TP e TA (Lawson 2001, p. 280), è necessario che al computer siano proposti precisi indicatori di equivalenza nei due testi, così da riconoscerne la corrispondenza e procedere all'allineamento (Koller 1995, Bowker & Pearson 2002, pp. 92-108). Tali indicatori sono determinabili secondo criteri linguistici e strutturali, ovvero in termini di singole parole o attraverso unità più complesse (frasi o enunciati, paragrafi, capitoli, e così via). Tuttavia, l'attività di traduzione non si limita certo alla mera trasposizione di equivalenze lessicali o sintagmatiche, così come non sempre lingue diverse presentano caratteristiche morfologiche e strutturali speculari. Se questo è stato confermato per i testi scritti, lo si trova ancora più accentuato nella traduzione della comunicazione parlata, dove la pressione temporale a cui sono sottoposti gli interpreti incide ulteriormente sulle possibilità di realizzazione del TA, indipendentemente dalla modalità considerata, con conseguenti dissimmetrie a livello lessicale e strutturale.[60] Ad ogni modo, nei corpora di lingua scritta l'allineamento sembrerebbe meno difficoltoso, perché basato su unità (*sentence*) meglio definite dalla presenza della punteggiatura (McEnery & Wilson 2001, p. 70), a differenza di quanto si avrebbe invece in un corpus di trascrizioni di testi orali.

A fronte di tutti questi ostacoli, Mikhailov (2001) individua due approcci generali all'allineamento intertestuale (riguardo alla traduzione scritta di testi narrativi), ossia un *realistic approach* e un *romantic approach*. Con l'approccio realistico si va alla ricerca di corrispondenze a livello strutturale, sulla base della constatazione che nei TA considerati è preservata al massimo la struttura in paragrafi dei TP, al pari del numero di parole presente in ogni testo. La misurazione di questi dati è effettuata per ogni paragrafo in modo da ottenere il cosiddetto *Source Language-Target Language quotient*. Questo valore è poi utilizzato per verificare la corrispondenza di ogni singolo paragrafo nel TP e nel TA: se il valore è simile a quello ottenuto nel calcolo generale per la coppia di lingue interessata, l'allineamento tra i due paragrafi viene eseguito; diversamente, se il valore si discosta eccessivamente, il sistema impiegato da Mikhailov richiama automaticamente il paragrafo successivo o precedente, a seconda che la compensazione sia necessaria nel TP o nel TA. I risultati ottenuti sembrano aver richiesto l'intervento diretto dell'analista solamente nel 10% dei casi. Tuttavia, i margini di miglioramento si amplificano se cambiano le lingue

[59] Addirittura Lawson (2001, pp. 293-294) afferma che «Parallel corpora of spoken language are unlikely, for instance, to be created because of the problem of cost and effort in the original language, and the almost impossible nature of translation of idiomatic spoken material into the target language».

[60] La presenza di dissimmetrie è tra l'altro confermata anche per coppie di lingue affini, quali lo spagnolo e l'italiano, ad esempio nella modalità simultanea (Russo 1990, 1997).

da allineare. Il secondo approccio, ovvero l'approccio romantico, si basa su un meccanismo simile al precedente, ma attinge non tanto dal calcolo del quoziente prima indicato, quanto da una «very sophisticated knowledge base» (*ibid*., p. 95). Questa comprenderebbe informazioni di varia natura sulle lingue e sui testi da allineare, similmente all'impostazione data al programma per allineare testi paralleli elaborato da Hofland & Johansson (1998). Il funzionamento di questo programma si basa su una *anchor list*, cioè una sorta di banca dati contenente un lessico bilingue e creata manualmente. A questa sono affiancate determinate regole, tra cui la presenza di lettere maiuscole all'inizio di frase o per i nomi propri, la presenza di segni di interpunzione o ancora le indicazioni su particolari formattazioni del testo (questi dati sono registrati in una lista a parte), e così via. L'allineamento è dunque operato attraverso l'inserimento automatico di un *tag* di inizio e di fine frase contenente un codice identificativo, il quale è aggiunto anche nel TA (o TP a seconda dei casi) corrispondente. Il grande vantaggio offerto da questo programma è che si è dimostrato in grado di allineare TP e TA con un alto tasso di precisione (è stata calcolata una percentuale di errore dell'1,98% su un corpus di 1,3 milioni di parole, con 51 testi e 93.000 frasi) in maniera automatica.

Per quanto riguarda i corpora orali paralleli, ad oggi non abbiamo trovato esempi simili di allineamento automatico. Purtroppo, al momento vi sarebbe una scarsa disponibilità di strumenti e solo per l'allineamento manuale. Due dei programmi precedentemente descritti (ADALaab e Exmaralda) sembrano essere gli unici a poter gestire TP e TA allineati e ad offrire funzionalità di ricerca automatica. In essi è rispettato l'allineamento tra TP e TA in base allo sviluppo temporale effettivo dei due flussi comunicativi, cioè in rapporto al *décalage* nell'interpretazione simultanea. In alternativa, vale la pena sottolineare che nei corpora di interpretazione simultanea si potrebbero allineare i TP con i TA anche sulla base del contenuto, trattando cioè il TP e il TA come due "testi autonomi" a livello di produzione temporale, cercando corrispondenze lessicali e strutturali al pari di quanto avviene nei corpora paralleli di lingua scritta.[61] In questo caso si dovrebbero sicuramente affrontare i problemi elencati prima in merito alle differenze lessicali e strutturali dei TP rispetto ai TA. Questo sarebbe vero per un allineamento di tipo automatico, ma volendo avvalersi di procedure tradizionali (allineamento manuale o semiassistito) si riuscirebbe ad ampliare lo spettro degli strumenti informatici a disposizione. Ad esempio, si potrebbero utilizzare quelli menzionati da Lawson (2001, pp. 290-292), come WordSmith (Scott 2003) e ParaConc (Barlow 2001/2003). Altri ancora sono stati esaminati da Zanettin (2001), per esempio Multiconcord[62] sviluppato da David Woolls

[61] Nell'ambito dei CTS, per un esempio di applicazione di alcuni dei programmi e degli approcci citati si veda Comastri (2002).
[62] Per un esempio di studio appartenente ai CTS con questo programma si veda Ulrych (1997).

(King & Woolls 1996, St.John & Chattle 1998) e Trados WinAlign. L'opzione migliore per un corpus di interpretazione sarebbe poter unificare le diverse metodologie, con operazioni automatiche e manuali (ma assistite) e, allo stesso tempo, soddisfare entrambe le prospettive di allineamento, sia quella intertestuale (per entrambi le opzioni, cioè sulla base del *décalage*, così come sulla base del contenuto), sia quella intratestuale (collegando la rappresentazione scritta del testo alla sua registrazione audio/video). Una tale complessità di configurazioni possibili si riflette inevitabilmente anche sulle modalità di accesso al corpus e sui diversi strumenti utilizzabili per studiarne i dati ed esplorarlo.

1.3.6 Accessibilità e distribuzione

Quasi tutte le diverse applicazioni informatiche a cui si è fatto riferimento nel discutere le tappe di trascrizione, codifica, annotazione e allineamento nella creazione di un corpus costituiscono strumenti con cui accedere ai materiali veri e propri raccolti in una tale risorsa, al fine di studiarli, ricavarne informazioni ed estrarne occorrenze rispetto a fenomeni che si intende esaminare. Questo tipo di utilizzo del corpus presuppone che il ricercatore abbia a disposizione tutti i file necessari (nello specifico, i file di testo con le trascrizioni e i file multimediali delle registrazioni, oppure i file XML o in altri formati delle trascrizioni annotate e codificate), così come sono stati impostati da chi ha realizzato il corpus. In alternativa, alcuni progetti predispongono l'accesso ai dati da parte della comunità scientifica attraverso vari tipi di interfaccia di ricerca *online*, dotate di una serie di funzioni e opzioni di ricerca. Da una interfaccia si ha dunque sempre accesso ai dati raccolti nel corpus, ma le possibilità di ricerca sono limitate alle funzioni e ai programmi di cui è corredata l'interfaccia stessa. Inoltre, non è detto che i risultati possano essere esportati se il sistema di accesso ne consente solo la visualizzazione.

In entrambi i casi, chi desideri accedere al corpus deve avere una certa familiarità con le principali nozioni di linguistica computazionale e con il funzionamento dell'applicazione in uso. Purtroppo, si deve riconoscere che il solo pensiero di utilizzare un nuovo programma informatico si presenta ai più come un'impresa estremamente impegnativa. Per questo, è importante che le modalità di accesso al corpus siano documentate nel modo più dettagliato ed efficace al tempo stesso. In questo modo, si dovrebbe riuscire a ridimensionare le difficoltà che si porrebbero nell'utilizzare strumenti messi a punto da altri ricercatori. Ammesso che sia così per quel che riguarda l'impiego degli strumenti per accedere a un corpus, la situazione non è altrettanto rosea se si considera la creazione vera e propria di nuovi strumenti e canali di accesso,

quali appunto interfaccia *online* o simili. In questo caso, è richiesto decisamente un apporto informatico che, il più delle volte, va ben oltre le competenze normalmente in possesso del *practiseracher* (Danielsson 2004). Diventa allora indispensabile profondere uno sforzo interdisciplinare e collaborativo, coinvolgendo e coordinando più soggetti che riescono a trovare un linguaggio di intesa e il giusto equilibrio tra ciò che è desiderabile e ciò che è effettivamente possibile.

Infine, l'accesso a un corpus potrebbe essere garantito attraverso la semplice distribuzione dei materiali che lo compongono (cioè i singoli file delle trascrizioni e delle registrazioni), assieme a tutta la documentazione pertinente.[63] Tale distribuzione può essere gestita dai creatori del corpus, oppure può essere affidata a istituzioni esterne, quali archivi o organizzazioni preposte alla raccolta e alla distribuzione di risorse linguistiche (ad esempio la *European Language Resources Association* e il *Linguistic Consortium*). In generale, l'accesso e la distribuzione di un corpus, ultima tappa perché se ne completi del tutto la piena realizzazione, pongono non poche questioni da affrontare con attenzione e cautela, al fine di assicurarsi che «for as long as possibile into the future, a corpus is useful and usable for a wide range of potential users» (Wynne 2005a). Lo stesso autore elenca diversi accorgimenti che si dovrebbero tenere in considerazione fin dall'impostazione iniziale del corpus, come a dire che ci si dovrebbe occupare sempre dell'ultima tappa a partire già dalla prima. I vari accorgimenti e suggerimenti possono essere così sintetizzati:

- mantenere una versione del corpus non annotata;
- porre un limite alle correzioni da apportare al corpus, stabilendo un livello di adeguatezza minimo, raggiunto il quale si può procedere al completamento del lavoro;
- assicurarsi che vi sia il rispetto dei diritti di proprietà intellettuale nei confronti di tutti i soggetti che hanno contribuito a vario titolo alla creazione del corpus;
- documentare tutti gli accordi presi ai fini della raccolta, archiviazione, analisi e distribuzione dei dati;
- provvedere a un adeguato backup dei dati;
- provvedere a una adeguata archiviazione[64] dei dati;
- garantire un accesso libero ai dati;

[63] Un esempio di questo tipo di distribuzione è dato dai corpora creati dai ricercatori che afferiscono al portale www.exmaralda.org, tra cui vi sono due corpora di interpretazione (Meyer 2008, Meyer & Schmidt s.d., §2.3).

[64] L'archiviazione è distinta dal backup, in quanto «Backup means taking a periodic copy of a file in store. Archiving means the transfer of information of public value into a separate repository where it is going to be held indefinitely, or for an agreed period of time» (*ibid.*).

- evitare l'uso di formati proprietari e prediligere l'impiego di formati aperti (per esempio XML e TXT);
- acquisire eventuali dati audio/video di alta qualità, operando compressioni o alleggerimenti solo su copie dei dati.

Dal numero di ostacoli e sfide metodologiche approfondite nel corso delle precedenti sezioni, non sorprende che il panorama di applicazione della Linguistica computazionale agli Studi sull'interpretazione sia molto più scarno rispetto a quanto è stato fatto negli Studi sulla traduzione. Ciononostante, è possibile tracciare un percorso di evoluzione ricco di esperienze interessanti, alcune delle quali pionieristiche, in questa branca emergente degli *Interpreting Studies*. Queste saranno illustrate nel prossimo capitolo, a cui seguirà la descrizione dettagliata del corpus EPIC e del corpus DIRSI-C, con le varie tappe che ne hanno scandito la realizzazione.

Capitolo 2

Albori e progressi dei CIS

2.1 Studi basati su corpora "manuali"

Nel tentativo di sondare a livello globale la disponibilità di *interpretation corpora*, Setton (s.d.) descrive una serie di progetti di ricerca appartenenti ai CIS (fino al 2003). Riporta ben 14 studi sull'interpretazione simultanea (di cui solamente due non comprendono la lingua inglese), tutti frutto di ricerche sul campo.[1] Tra questi, alcuni riguardano ricerche effettuate più di venticinque anni fa. Conseguentemente, le relative registrazioni e trascrizioni non sono sempre disponibili. In realtà, anche considerando i lavori più recenti, va ammesso che «frustratingly, it has rarely been possible to listen to other researchers' original tapes» (Setton 2002, p. 33). Pertanto, se l'accesso diretto alle trascrizioni può essere più facilmente garantito attraverso supporti cartacei e informatici, lo stesso non si può dire dei dati orali registrati (utilizzando, per esempio, interfaccia web) – un limite, questo, che trova conferma nella seguente affermazione: «publishers of interpreting corpora have no option but to exhort readers to 'read with their ears'» (*ibid.*). Inoltre, per quanto riguarda le dimensioni dei corpora presi in esame, si andrebbe da un minimo di sette / trenta minuti a un massimo di 14 ore di dati registrati (nella maggior parte dei casi, le trascrizioni effettivamente prodotte ai fini dell'analisi sono una parte più limitata). Setton riassume le informazioni sui vari progetti in una tabella, di cui riproduciamo solo una parte (le prime tre colonne con le indicazioni di chi conduce la ricerca, su quali lingue e in che ambito):

[1] Si ha l'impressione che la distinzione iniziale tra gli studi basati su corpora e gli altri si riferisse prevalentemente al tipo di dati raccolti, cioè su base empirica-osservazionale (da situazioni reali) o su base sperimentale rispettivamente.

Figura 2.1 Studi CIS sondati da Setton (s.d.).

Researcher	**Languages**	**Event**
Oléron & Nanpon 1965	EN, FR, DE, ES	UNESCO impromptu (non-tech discussion)
Déjean Le Feal 1978	FR>DE	various speeches
Chernov 1979, forthcoming	EN, FR, ES, RU	UN 1968
	EN>RU, ES, FR	1978 UN satellite interpreting experiment
Lederer 1981	DE > FR	Railway Consortium and lab (2nd versions)
Shlesinger 1989	HEB><EN	Courtroom testimony
Donovan 1994	FR><EN	extracts 2 meetings 1986-8
Pöchhacker 1994	EN><DE, FR >DE	Vienna small business conference *ICSB*
Kalina 1998	EN><DE<>FR	*Bertell*1989 public lecture (anti nuclear)
	DE, FR, EN	*Würzburg* 1992 law symposium
Setton 1997, 99	DE>EN	extracts from Kalina *Würzburg* corpus
Wallmach 2000 & forthcoming	EN, Zulu, Afrikaans, Sepedi	Parliament speeches Gauteng (S. Africa) Provincial legislature
Diriker 2001	EN><TR	Conference on Metaphysics and Politics (2 days
Vuorikoski 2004	mostly DE, EN ><FI, SW	European Parliament debates
Beaton, forthcoming		European Parliament debates via EBS
Monacelli, forthcoming	IT><EN	10 speeches from 4 events conferences

Si rende necessario puntualizzare che fino a questo punto si è sempre utilizzato il termine "corpus", nonostante gli studi citati da Setton fossero basati su un tipo di analisi che non prevedeva quasi mai l'ausilio di strumenti appartenenti alla linguistica computazionale. Egli stesso parla infatti di «'manual' corpus studies» (*ibid.*), probabilmente intendendo con questo che le trascrizioni erano analizzate in formato cartaceo (su stampa) e che le registrazioni erano state effettuate su nastro in situazioni reali. Alcuni esempi particolarmente interessanti sono le ricerche di Pöchhacker (1994b) e Kalina (1998), i quali hanno utilizzato materiali registrati da convegni reali e non da prove effettuate in contesti simulati o in laboratorio; lo stesso Setton (1999) ha utilizzato una parte del corpus di Kalina, a cui ha aggiunto una parte ulteriore di materiale ottenuto in condizioni sperimentali. Lo studio di Pöchhacker è citato anche da Riccardi (2009), assieme ai lavori di Lederer (1981), Mackintosh (1983) e Di Guida (2001), non tanto in riferimento al *corpus-based approach*, quanto per l'aver esaminato eventi comunicativi riconducibili alla conferenza-convegno (mediati

da interpreti) nella loro interezza.[2] Tuttavia, «si lamenta ancora la scarsità di indagini e rilevazioni dedicate agli 'eventi comunicativi con IS' [interpretazione simultanea]» (Riccardi 2009, p. 362). Questa affermazione ci spinge a porci il seguente interrogativo: qual è la situazione dei CIS nel nuovo millennio, dopo ben oltre dieci anni di ricerche dall'appello lanciato da Shlesinger (1998), nel corso dei quali lo sviluppo delle tecnologie dell'informazione ha migliorato notevolmente le possibilità di trasmissione e registrazione dei dati? Per quanto vi sia ancora tanta strada da percorrere, il panorama generale presenta un certo numero di esperienze che potrebbero ispirare futuri progetti di ricerca.

Nell'ambito di uno studio sempre di tipo "manuale", ma svolto su un campione di dimensioni considerevoli, in parte reso anche disponibile in formato elettronico, Vourikoski (2004)[3] ha creato e analizzato un corpus quadrilingue (inglese, finlandese, svedese e tedesco) per un totale di 122 discorsi registrati nel contesto dei dibattiti al Parlamento europeo. Ogni testo di partenza è accompagnato dalla relativa interpretazione simultanea in tutte le combinazioni possibili tra le quattro lingue selezionate; in questo caso, è interessante notare che le registrazioni di alcuni esempi di interventi sono disponibili in formato digitale e collegati direttamente al file della trascrizione (in formato elettronico). La trascrizione è strutturata secondo uno schema tabulare, con il TP distribuito lungo una colonna, affiancata dai tre TA corrispondenti. L'uso di schemi tabulari anche in fase di analisi si è rivelato estremamente utile al fine di gestire una tale struttura del corpus, potendo infatti

[2] Lo studio di Galli (1990) su un campione di presentazioni registrate da convegni di ambito medico è estremamente pertinente con il tipo di dati raccolti in DIRSI, ma non può essere incluso tra questi contributi che si sono occupati della situazione comunicativa globalmente. Galli ha infatti considerato esclusivamente le relazioni o comunicazioni (eventi linguistici a cui si riferisce sempre con i termini *speeches* e *texts*) presentate da dieci conferenzieri, con il coinvolgimento di tre interpreti simultanei professionisti. Il campione comprende quattro interventi in italiano (per un totale di 7.000 parole) e sei in inglese (solo due da madrelingua, per un totale di 15.500 parole). La durata dei vari interventi si aggira intorno ai 20 minuti (con una variazione che spazia da un minimo di 8 a un massimo di 25 minuti). Le registrazioni sono state trascritte ortograficamente, ma non letteralmente in quanto «pronunciation mistakes, pauses and non linguistic vocalization were not [transcribed]» (*ibid*., p. 64). Pur trattandosi di un'analisi manuale, essa è stata svolta sulla base di un'annotazione particolare, con la quale sono stati segnalati fenomeni che rientrano nella categoria "*departures*": omissioni, aggiunte, sostituzioni, errori di traduzione e interpretazioni (quest'ultimo tipo comprende tutto ciò che non rientra nelle precedenti categorie), differenziandone vari tipi in ciascuna categoria. Infine, i risultati sono stati verificati rispetto ai parametri velocità di eloquio, direzionalità e modalità di presentazione del TP (*prepared* vs. *semi-prepared*). Va specificato che i TA analizzati in questo studio non sono stati registrati durante lo svolgimento di un convegno vero e proprio. Le registrazioni dei TP provengono da una convegno reale e sono state riproposte ai tre interpreti in un secondo momento, seguendo le stesse procedure del convegno reale da cui erano state ottenute. È questa un'interessante alternativa empirica che si collocca a metà strada tra l'analisi osservazionale (in contesti reali) e l'analisi sperimentale (in contesti simulati), ovvero una «verifica sperimentale di dati empirici» (Riccardi 2009, p. 359).

[3] La stessa Vuorikoski (2004, pp. 32-41) offre un interessante approfondimento degli studi condotti su quelli che chiama «real-life corpuses», citando Lederer (1981), Pöchhacker (1994), Kalina (1998) e Setton (1999).

annotare (per quanto questa operazione sia stata effettuata manualmente) la presenza o meno dei fenomeni studiati nelle tre diverse rese dei TA in parallelo rispetto allo stesso TP. In definitiva, nonostante l'analisi condotta sia ancora di tipo tradizionale, questi materiali potrebbero essere facilmente predisposti per la realizzazione di un corpus elettronico a tutti gli effetti.

Un altro contributo degno di nota è dato dall'immenso campione di dati studiato da Straniero Sergio (2007), concernente una vasta gamma di trasmissioni televisive con la presenza del servizio di interpretazione simultanea e consecutiva. Il campione analizzato comprende 943 estratti con la presenza di 107 diversi interpreti in 104 programmi, tutti trasmessi alla TV italiana lungo un arco di tempo di circa trenta anni. Tale campione fa parte di una raccolta di dimensioni maggiori e che andrà a costituire il CorIT – Corpus di Interpretazione Televisiva. Questo corpus in fase di realizzazione dovrebbe comprendere un totale di 2.340 interpretazioni (suddivise tra talk show, eventi mediali e conferenze stampa del Gran Premio di Formula 1).

Nella parte di analisi condotta da Straniero Sergio è stato applicato l'approccio dell'Analisi conversazionale, con l'adozione quindi delle relative convenzioni per la trascrizione dei dati e l'approfondimento delle «molteplici articolazioni del talk show in quanto conversazione-spettacolo» (*ibid*., p. 21). Il tipo di approccio e le conseguenti convenzioni di trascrizione adottate potrebbero rendere difficoltosa la lettura semiautomatica dei dati attraverso programmi di linguistica computazionale. Pertanto, in questo caso sarebbe forse necessario uno sforzo maggiore per strutturare i materiali in un corpus elettronico, sforzo che sembrerebbe essere in corso di attuazione attraverso l'uso del software Winpitch (Martin 2004).

Il parlato dialogico mediato da interpreti è l'oggetto di studio di un altro corpus in fase di realizzazione, chiamato *Dialogue Interpreting Corpus* (Merlini 2007, pp. 286-287). I dati in esso raccolti provengono sia da situazioni reali, sia da situazioni simulate in ambito didattico, e riguardano diverse aree, quali i servizi per i migranti, i servizi sociosanitari e le trattative commerciali. Tutte le trascrizioni sono già state completate e, stando agli esempi riportati dall'autrice (*ibid.*), contengono annotazioni conformi alle convenzioni utilizzate nell'Analisi della conversazione.[4]

Tornando all'interpretazione simultanea fornita in occasione di conferenze e convegni, Ahrens (2004, 2005) ha svolto un'analisi del profilo prosodico della resa di sei interpreti che, distribuiti in tre cabine parallele, avevano tradotto in simultanea (inglese > tedesco) lo stesso TP della durata di

[4] Niemants (2009) sta raccogliendo un campione di dati simile in ambito sociosanitario, comprendente cioè interazioni reali e anche simulate in sede di esame universitario, con l'obiettivo di creare un corpus elettronico (progetto di tesi di dottorato in corso presso l'Università degli Studi di Morena e Reggio Emilia. La descrizione del progetto è disponibile *online* all'indirizzo <http://www.dailynterpreter.com/my-writings>).

72 minuti. L'evento comunicativo in questione è una conferenza reale, tenuta all'università, per la quale era stato possibile ingaggiare contemporaneamente tre équipe di interpreti professionisti. Questi ultimi, oltre ad essere registrati, hanno anche partecipato allo studio compilando un questionario al termine dell'incarico. Le registrazioni sono state realizzate su doppia pista, digitalizzate e trascritte, in modo da poter studiare i dati con un software apposito (PRAAT) che consentisse di esaminare alcune caratteristiche prosodiche fondamentali, tra cui le pause, le unità tonali, i contorni intonativi e l'altezza. Pur riconoscendo che in questo contributo il corpus sia stato studiato attraverso una «computer-aided analysis» (Ahrens, 2005, p. 57), non sembrano esserci esplicite indicazioni su metodi di estrazione semiautomatici delle occorrenze dei fenomeni oggetto di studio, così come non vi sono indicazioni sull'eventuale disponibilità diretta del corpus. Per questi motivi, il corpus in questione rientra ancora nel tipo "tradizionale", per quanto vi siano tutti i presupposti perché si possa utilizzare il materiale per la realizzazione di un corpus elettronico vero e proprio.

Anche Diriker (2004) ha basato la sua analisi di fenomeni pragmalinguistici che segnalano «how simultaneous interpreters are "positioned" in an actual conference context» (*ibid.*, p. 50) utilizzando un corpus, o meglio un campione, tratto da un convegno di due giorni svoltosi in un contesto universitario (per la precisione, si tratta del Dipartimento di Filosofia dell'Università Bogaziçi a Istanbul). Per tale evento della durata di due giorni sono stati ingaggiati tre interpreti, due per la cabina inglese e uno per la cabina francese. Quest'ultimo è stato ingaggiato solo per la mattina del secondo giorno e per un singolo intervento in francese, il quale è stato tradotto solo in turco, senza quindi che fosse fornita anche la resa in inglese attraverso il *relais* tra le due cabine. Tutto il materiale è stato registrato su cassetta, con non pochi inconvenienti tecnici e pratici, ed è stato poi trascritto ortograficamente, secondo convenzioni estremamente semplici e mirate a ottenere un testo trascritto che fosse facilmente leggibile da parte dell'analista. Per questo motivo, è stata indicata una minima parte dei tratti paralinguistici, quali l'intonazione (inserendo segni di interpunzione) e le pause (vuote e piene). È curioso constatare che la dimensione totale del campione sia fornita indicando non tanto le ore di registrazione o il numero di parole o sillabe, bensì il numero di pagine trascritte (circa 120), analizzate per l'appunto manualmente. In questo caso, tutti i dati raccolti (non solo TP e TA, ma anche interviste ai partecipanti e appunti sulle dinamiche di svolgimento del convegno) potrebbero essere strutturati in un corpus multimodale,[5] in modo da ottenere uno sguardo d'insieme sulle tante sfaccettature rilevate da Diriker nella situazione comunicativa del convegno.

[5] Per un esempio di trascrizione multimodale di una lezione frontale universitaria (*lecture*) si veda Crawford Camiciottoli (2004, p. 51).

Infine, un altro esempio è lo studio di Monacelli (2009), la quale ha analizzato un corpus contenente le trascrizioni di dieci interventi (e della loro interpretazione simultanea) nell'ambito di alcuni convegni che si sono svolti in Italia.[6] Nell'insieme, sono stati coinvolti dieci interpreti professionisti, di cui cinque sono membri di AIIC, l'Associazione Internazionale Interpreti di Conferenza. In totale il corpus arriva a 119 minuti di materiale originale (testi di partenza) così suddiviso: tre testi per la combinazione francese > italiano; sei testi per la combinazione inglese > italiano; un testo per la combinazione italiano > inglese. La durata varia dai 5 ai 35 minuti circa e la direzionalità degli interpeti è sempre da B a A, lavorano cioè tutti verso la loro lingua materna. Anche in questo caso le trascrizioni sono state strutturate secondo uno schema tabulare per poi essere analizzate "manualmente", probabilmente anche a causa dei fenomeni presi in esame. L'attenzione è stata posta soprattutto su alcuni aspetti sociolinguistici e interazionali finora poco studiati, quali il *participation framework* e la *interactional politeness* degli interpreti partecipanti alla situazione comunicativa.

Se si considera la definizione di corpus che abbiamo dato all'inizio del primo capitolo, in tutti i progetti presentati in questa sezione più che di "corpus" si potrebbe parlare di "banca dati", "campione" o "campionatura". Ciononostante, è un dato di fatto che in letteratura si riscontra quasi sempre il riferimento a "corpus" anche laddove i dati non sono strutturati in formato elettronico e in modo tale da poter essere analizzati con strumenti informatici, o resi accessibili ad altri ricercatori. In realtà, negli ultimi anni sono stati realizzati esempi di *machine-readable corpora* anche nel campo dell'interpretazione, applicando appieno quindi la linguistica computazionale allo studio della traduzione della comunicazione parlata.

2.2 Studi basati su corpora "elettronici"

Uno dei primi esempi di corpora effettivamente leggibili dal computer e realizzati nel campo degli Studi sull'interpretazione è il TIC – *Television*

[6] Uno degli aspetti metodologici più interessanti di questo studio è che la ricercatrice ha avuto l'opportunità di reperire registrazioni di eventi che erano stati tenuti prima ancora di stabilire l'oggetto di studio vero e proprio. In questo modo, ha potuto coinvolgere in più fasi della ricerca i soggetti interessati (cioè gli interpreti) senza alterare i dati o i risultati. Dopo aver ottenuto i dati, infatti, organizza prima una sessione di briefing con intervista (al fine di tracciare un profilo completo dei soggetti coivolti), poi svolge l'analisi testuale e infine tiene una sessione di de-briefing. Con questa scelta è stato possibile «both to foster the active participation of the subject and to maintain the rigor required so as not to taint the data with the analyst's personal comments» (Monacelli 2009, p. 29).

Interpreting Corpus messo a punto da Cencini (2000) per la sua tesi di laurea. Nel TIC sono raccolte le trascrizioni di 11 testi tratti da eventi mediali trasmessi alla televisione e mediati da interpreti tra le lingue inglese e italiano, per un totale di circa 36.000 parole e quattro ore di registrazione. Nonostante le dimensioni esigue, il TIC è un primo esempio significativo all'interno dei CIS, in quanto ha per la prima volta affrontato diverse questioni fondamentali nell'applicazione della Linguistica dei corpora agli Studi sull'interpretazione. In effetti, il TIC è presentato come un «corpus pilota e il suo scopo principale è quello di proporre un sistema per adeguare lo schema definito dalla TEI alle esigenze di codifica (e di conseguenza analitiche) dei testi di interpretazione» (*ibid*., p. 60). Il software utilizzato da Cencini per svolgere alcune analisi sui dati del corpus è SARA – *SGML Aware Retrieval Application* (Aston & Burnard 1998), oggi conosciuto come XAIRA nella sua versione aggiornata (cfr. sitografia). I parametri di codifica definiti nel TIC sono un punto di partenza prezioso nella definizione di elementi e attributi da utilizzare all'interno di un corpus per l'interpretazione. Essi sono inseriti sotto forma di *tag* che rispondono al formato XML sia all'interno del testo delle trascrizioni, sia nell'intestazione di ciascuna trascrizione. Tale intestazione (*header*) raccoglie tutta una serie di informazioni metatestuali, quali il contesto di provenienza della trascrizione/registrazione, la modalità di interpretazione, la funzione dell'interprete, il nome del trascrittore, i partecipanti all'interazione, le loro caratteristiche e così via (Cencini & Aston 2002, §1.3.4). Nonostante vi sia la possibilità di visualizzare una trascrizione codificata secondo lo standard TEI in modo *user-friendly*, la grande quantità di attributi considerati nel TIC (tra cui molti tratti paralinguistici) e l'ampiezza dell'approccio adottato potrebbero rendere difficoltoso il suo impiego da parte di chi ha poca dimestichezza con i linguaggi informatici. Si potrebbe pensare cioè a un primo livello di strutturazione più semplice, soprattutto in termini di attributi da includere nelle trascrizioni, da adottare poi come base "grezza" a cui aggiungere ulteriori informazioni attraverso elaborazioni più complesse. Pur ammettendo infatti che «we want our transcription to be *machine-friendly*» (*ibid*., p. 51), di modo che possa essere elaborata dal computer, in tutto questo sarebbe comunque auspicabile poter tutelare anche la dimensione "*user-friendly*" a favore degli utenti (per non dire "*transcriber-friendly*" o "*annotator-friendly*" in considerazione di chi deve produrre le trascrizioni).

Un contesto comunicativo completamente diverso è stato preso in esame da Wallmach (2002b), la quale ha analizzato l'interpretazione simultanea fornita durante alcune sedute del principale organo giuridico nella provincia sudafricana del Gauteng. In questa area del paese sono comunemente parlate quattro delle undici lingue ufficiali del Sud Africa, per cui durante le sedute del *Gauteng Provincial Legislature* è previsto un servizio di interpretazione simultanea da tutte le lingue ufficiali verso l'inglese, l'Afrikaans, lo Zulu e lo Sepedi. Il

contesto lavorativo è particolarmente complesso, al pari di quello sociocomunicativo nel suo insieme. Il primo è caratterizzato da testi altamente tecnici, preparati e letti ad alta velocità; rispetto al secondo, tutte le lingue in questione hanno una diffusione limitata nella geografia del paese e, in molti casi, non dispongono di terminologia giuridica e tecnica adeguatamente sviluppata. A questo si aggiunga che l'interpretazione simultanea è un servizio entrato solo in tempi recenti a far parte della "vita comunicativa" sudafricana (Walmach 2002a, 2004).[7] A fronte di un siffatto panorama, con il suo studio Walmach si è proposta di analizzare le caratteristiche dei TA, costruendo un corpus contenente le trascrizioni tratte da circa quattro ore di registrazioni (in Afrikaans e Zulu) eseguite su cassetta con una doppia pista. I TP e i TA sono stati allineati manualmente a livello di frase, utilizzando il programma ParaConc.[8] In questo modo, è stato possibile ottenere un riscontro immediato della gestione delle sfide poste dalla tecnicità e dalla velocità dei TP da parte degli interpreti simultaneisti coinvolti (due per ogni combinazione linguistica). Pur ammettendo che il corpus «needs to be developed much further before it can be successfully exploited» (*ibid*., p. 509), questo studio dimostra la grande potenzialità dei CIS nei contesti caratterizzati da un alto grado di multilinguismo.

Un altro corpus contenente TA prodotti in condizioni sperimentali è stato creato da Timarová (2005), con il coinvolgimento di 18 soggetti tra interpreti studenti e neolaureati in interpretazione. Ai partecipanti è stato chiesto di fornire l'interpretazione dall'inglese in ceco (dalla lingua C alla lingua A) di due discorsi registrati, di cui uno da rendere in modalità simultanea e uno in modalità consecutiva. Le trascrizioni ottenute dalle registrazioni di tutte le prove hanno dato vita a un corpus di 30.000 parole, analizzato grazie all'uso dei programmi WordSmith Tools e MS Excel. Gli esempi di analisi riportati (uno sulla lunghezza dei testi, misurata contando sia il numero di parole sia il numero di sillabe, e uno sulla densità lessicale) evidenziano quanto sia stato vantaggioso (per non dire indispensabile) l'uso di programmi di linguistica computazionale, poiché «Many previously laborious steps in data analysis can be done as,

[7] Lo sviluppo di tutti i servizi di Traduzione in generale ha seguito un corso peculiare in questo paese. I motivi sono assai diversi, tra cui il grado di diffusione delle lingue divenute ufficiali dopo la fine dell'Apartheid e il "regime socio comunicativo" imposto dagli stessi sostenitori di questa politica di segregazione razziale (Kruger 2008).

[8] Un corpus di interpretazione consecutiva è stato creato da Fumagalli (1999/2000), utilizzando un programma analogo (MultiConcord) per l'allineamento. Tale corpus include 18 TA interpretati dall'inglese in italiano, in modalità consecutiva, da studenti di interpretazione; questo è stato raffrontato anche con un altro corpus di 15 discorsi originali in italiano per un'analisi comparabile. Obiettivo di questo lavoro è stato esplorare la presenza di caratteristiche tipiche dei testi tradotti, trovare cioè traccia di ciò che è conosciuto come *translationese* (esplicitazione, semplificazione, normalizzazione e livellamento, si veda Baker 1996), effettuando un confronto con le caratteristiche non solo dei TP corrispondenti, ma anche con quelle di testi prodotti originariamente in italiano (la stessa lingua dei TA analizzati). Un altro apporto pionieristico nell'ambito dei corpora di interpretazione consecutiva è di Dollerup e Ceelen (1996).

literally, one-click operations on a large number of data files» (Timarová 2005, p. 65). Ad esempio, nello studio sulla lunghezza dei TA in ceco è stato possibile ottenere il numero di sillabe estraendo automaticamente l'occorrenza delle vocali e dei dittonghi presenti nel corpus; dall'altra parte, nello studio sulla densità lessicale sono state prodotte automaticamente le liste di frequenza di tutti i lemmi. Tuttavia, l'assenza di una annotazione grammaticale come il *POS-tagging* non ha consentito di separare automaticamente le *content words* dalle *function words*: «This step had to be done manually by going through the list of 3967 words» (*ibid*., pp. 68-69). Ad ogni modo, è questo un altro esempio di come i CIS siano in grado di contribuire non solo alla ricerca di tipo quantitativo, ma anche qualitativo.

Tra gli esempi di corpora elettronici e analizzati con metodi appartenenti alla *Corpus Linguistics* va annoverato anche un corpus comparabile intermodale in quanto «consisting solely of translations, in different modalities or in different modes» (Shlesinger 2008, p. 240). Dopo uno sforzo considerevole di progettazione del corpus e dello studio stesso, Shlesinger è riuscita a comporre un corpus con molteplici TA (ottenuti a partire dallo stesso TP), prima come traduzione scritta e, successivamente, a distanza di oltre tre anni, come resa in interpretazione simultanea, coinvolgendo sempre gli stessi sei soggetti (traduttori e interpreti professionisti) dall'inglese (lingua B) in ebraico (lingua A). L'analisi si è concentrata su elementi tipicamente affrontati nella *Corpus Linguistics*, quali la varietà lessicale, calcolata sulla base della *type-token ratio*, e altre caratteristiche lessicali e grammaticali della lingua ebraica (per esempio, l'uso dei sistemi verbali, l'uso dell'articolo definito, la distribuzione delle POS – *Part-of-Speech* e così via), con l'obiettivo di individuare tratti salienti del cosiddetto interpretese. A detta della stessa autrice, «The methodology adopted in the present study would not have been possible without MorphTagger, a sophisticated morphological analyzer, or else it would have been limited to manual counts and intuitive judgments» (*ibid*., p. 244).

Infine, l'unico esempio pionieristico di corpus multimodale riscontrato ad oggi nei CIS è un piccolo corpus per lo studio dell'interpretazione tra l'inglese americano e la lingua dei segni italiana o LIS (Kellet Bidoli 2004, 2007).[9] Le criticità poste dalla rappresentazione elettronica, in un'unica configurazione, della comunicazione parlata e del linguaggio dei segni trovano finalmente una soluzione nella tipologia di corpus multimodale. Grazie all'impiego di un software apposito (C-I-SAID – *Code-A-Text Integrated System for the Analysis of Interviews and Dialogues*), è stato possibile interfacciare la trascrizione ortografica di quattro comunicazioni, presentate nell'ambito di convegni su temi

[9] L'unico esempio antecedente a questo che abbiamo potuto rilevare è lo studio di Cokely (1992). Tuttavia, in questo studio non era stato predisposto un corpus elettronico vero e proprio ed erano state selezionate solamente alcune parti a campione del materiale registrato (20% dei 200 minuti totali).

inerenti alla linguistica, con la trascrizione corredata da glosse dei rispettivi TA prodotti in LIS. L'operazione ha richiesto uno sforzo titanico, basti considerare che «The parallel transcription process took four months to complete owing to the trilingual nature of the study and the three-dimensional form of signs as well as gestures that often produce meaning not easy to gloss» (Kellet Bidoli, 2007, p. 335). Le trascrizioni sono poi state impostate secondo un formato a spartito e segmentate in unità definite dall'andamento prosodico e gestuale dei codici registrati. Non è chiaro se per le prime analisi ci si sia avvalso di estrazioni automatiche di alcune occorrenze, ma il programma utilizzato sembrerebbe consentire lo svolgimento di questo tipo di ricerche, con inclusa la possibilità di esportare i risultati in fogli di lavoro adatti all'analisi statistica. È evidente che l'applicazione del *corpus-based approach* allo studio dell'interpretazione in lingua dei segni richiede un impegno ancora maggiore a favore della collaborazione interdisciplinare, dovendo fare obbligatoriamente affidamento ad applicazioni sempre più avanzate rispetto alla gestione di dati disponibili in formati molto diversi tra loro.

2.3 Studi basati su corpora "elettronici" e pubblicamente accessibili

Nonostante i corpora descritti nella sezione precedente siano a tutti gli effetti corpora elettronici, i materiali in essi contenuti non sono ancora accessibili direttamente dall'esterno, per esempio attraverso un'interfaccia web. Sono ancora pochi i progetti che dispongono anche di questa ulteriore caratteristica, o che hanno raggiunto una fase di completamento del corpus per cui questo viene distribuito su richiesta degli interessati.

Uno di questi pochi esempi è il SIDB – *Simultaneous Interpretation Database*, sviluppato a partire dall'inizio dell'anno 2000 presso il CIAIR – *Center for Integrated Acoustic Information Research* della *Nagoya University* in Giappone. Stando alla documentazione che abbiamo consultato (Matsubara et al. 2002, Tagaki et al. 2002, Tohyama et al. 2005, Ono et al. 2008), sembra che uno dei principali obiettivi iniziali di questo progetto riguardasse lo sviluppo di tecnologie volte a creare un sistema di traduzione simultanea automatizzato. Tuttavia, e fortunatamente, questo è solo uno degli obiettivi, tra i quali rientra anche il miglioramento della formazione degli interpreti simultaneisti (Tohyama & Matsubara 2006). Di certo si tratta di un progetto dalle caratteristiche assai interessanti, a partire dalla dimensione: quasi un milione di parole tra TP in inglese e giapponese (comprendenti conferenze in formato monologico e interazioni simulate in formato dialogico) e i rispettivi TA, per un totale di oltre 180 ore di registrazione. Le rese in simultanea sembrano essere state effettuate solo per le conferenze, con il coinvolgimento di 21 interpreti professionisti in

totale, tutti madrelingua giapponese. Lo stesso TP è sempre stato interpretato da due o da quattro interpreti contemporaneamente, così da avere a disposizione molteplici TA dello stesso TP e poter esaminare diversi approcci. Tuttavia, pare che siano stati registrati solo i primi dieci minuti di ciascuna conferenza, così come sembrerebbe che tutte le conferenze siano state organizzate appositamente ai fini della realizzazione del corpus in un contesto simulato. Le trascrizioni di questo corpus parallelo sono anche allineate sulla base dell'enunciato; quest'ultimo sarebbe definito come unità di segmentazione prendendo in considerazione le pause maggiori o pari a 200 ms, oppure del valore di 50 ms al termine di una frase. Tra gli studi effettuati sul materiale contenuto nel corpus SIBD, uno di questi (Tagaki et al. 2002) ha evidenziato un risultato interessante in merito al numero di enunciati prodotti dagli interpreti rispetto al TP corrispondente: nella direzione inglese > giapponese tale numero è risultato essere sei volte superiore a quello calcolato per i TP in inglese, così come nella direzione opposta (giapponese > inglese) tale numero è risultato essere cinque volte superiore a quello ottenuto nei TP giapponesi. Questo proverebbe che è stata effettuata una segmentazione maggiore del TA, dovuta alla rielaborazione che l'interprete compie sulla struttura del TP, al fine di migliorarne la trasmissione e la ricezione finale. Tuttavia, considerando la coppia di lingue in questione, il significato dei risultati ottenuti andrebbe sicuramente letto anche alla luce delle notevoli differenze esistenti tra i due codici.

In ambito europeo, uno dei pochi esempi di corpora elettronici i cui dati sono disponibili *online* è il corpus CoSi (Meyer 2008, Meyer & Schmidt s.d.). Si tratta di un corpus sull'interpretazione simultanea e consecutiva, contenente le trascrizioni (35.000 parole) ottenute da oltre cinque ore di registrazioni di TP in portoghese brasiliano e dei relativi TA in tedesco. I TP riguardano la stessa comunicazione (*lecture*) tenuta in tre diverse occasioni dalla stessa persona, la quale era stata invitata in Germania da una ONG impegnata nella tutela dell'ambiente. I TA sono la resa in consecutiva (nelle conferenze di Berlino e Amburgo) e in simultanea (nella conferenza di Heidelberg) fornita da un campione di cinque interpreti professionisti. In occasione della conferenza con il servizio di interpretazione simultanea sono state organizzate due cabine contemporaneamente (analogamente a quanto descritto nello studio di Ahrens 2004, 2005), entrambe per la stessa combinazione linguistica e con due interpreti (uno dei quali è lo stesso interprete ingaggiato per la consecutiva ad Amburgo). Tutti gli interpreti coinvolti sono madrelingua tedeschi e hanno pertanto lavorato nella direzionalità B > A.

Il corpus CoSi (denominato anche "K6" nel portale dedicato al progetto) è stato creato con il programma EXMARaLDA – *Extensible Markup Language for Discourse Annotation*, con il quale è possibile strutturare e visualizzare diversi livelli di trascrizione come in un pentagramma o in uno spartito, agganciare i singoli livelli alla traccia audio corrispondente, nonché utilizzare

ciascun livello per un particolare tipo di annotazione. Il programma per svolgere le ricerche semiautomatiche si chiama invece EXAKT – *EXMARaLDA Analysis and Concordancing Tool*. Entrambi i programmi sono disponibili *online* alla pagina web dedicata a questo promettente progetto di ricerca.

Tra le risorse sviluppate con questi software, vi è un altro corpus di interpretazione, chiamato DiK corpus (denominato anche "K2") e realizzato nell'ambito del progetto *Interpreting in Hospitals*. I materiali del corpus DiK provengono da trascrizioni di eventi linguistici monologici e di interazioni medico-paziente mediate da interpreti in ambito ospedaliero. Le lingue interessate sono varie, tra cui tedesco, turco, portoghese e spagnolo, per un totale di 25 ore di registrazione.

Infine, nella terza categoria di corpora di interpretazione possono essere annoverate le due risorse linguistiche oggetto del presente lavoro, e che saranno descritte in dettaglio nei seguenti capitoli. Il primo progetto che sarà presentato riguarda il corpus chiamato EPIC – *European Parliament Interpreting Corpus* (Bendazzoli et al. 2004, Monti et al. 2005, Bendazzoli & Sandrelli 2005/2007), un corpus trilingue di discorsi registrati dalle sedute plenarie del Parlamento europeo. Il corpus comprende sia discorsi originali in italiano, inglese e spagnolo, sia i rispettivi TA tradotti in simultanea in tutte le direzioni possibili tra le tre lingue coinvolte. Il secondo progetto, realizzato sulla base dell'esperienza acquisita con EPIC, riguarda il corpus DIRSI-C – *Directionality in Simultaneous Interpreting Corpus*, un corpus bilingue (italiano e inglese) di eventi linguistici registrati nel corso di tre convegni medici che hanno avuto luogo nel mercato privato italiano. Gli interpreti coinvolti (quattro madrelingua italiani e uno madrelingua inglese) hanno tradotto da e verso la lingua straniera, secondo la normale prassi riscontrabile in questo tipo di incarichi professionali. Per questo motivo, il fattore direzionalità è stato messo in primo piano già nel nome di questa nuova risorsa linguistica per la ricerca e la didattica dell'interpretazione.

Il presente lavoro e il corpus DIRSI-C stesso sono ampiamente basati sulla metodologia sviluppata nell'ambito del progetto di ricerca in cui EPIC è stato creato. Le questioni metodologiche e le sfide affrontate in questi due progetti CIS sono pressoché simili, per quanto vi siano anche ovvie differenze dovute alla diversità dei due contesti di provenienza dei materiali inclusi nei due corpora (il Parlamento europeo in EPIC e il mercato italiano dei convegni internazionali in DIRSI-C).

La panoramica delle ricerche presentate in questo capitolo, a partire dai primi studi svolti su dati estrapolati da situazioni reali ma con metodi ancora tradizionali, fino ai recenti corpora elettronici disponibili *online*, mostra chiaramente che si è verificata una notevole evoluzione all'interno dei CIS. In particolare, negli ultimi anni il *corpus-based approach* è stato finalmente

applicato nella sua interezza, portando alla creazione di risorse di enorme valore e a disposizione dell'intera comunità scientifica. Si tratta dunque di uno sviluppo piuttosto recente degli *Interpreting Studies*, rallentato purtroppo dalle svariate sfide e ostacoli metodologici discussi in precedenza. Gli stessi saranno ripresi nei capitoli successivi, nei quali è descritto dettagliatamente l'intero processo di creazione di EPIC e di DIRSI-C, approfondendo di volta in volta le questioni più significative nelle varie fasi della loro realizzazione.

Capitolo 3

L'Archivio Multimediale e il Corpus EPIC

La creazione di EPIC rappresenta uno dei primi tentativi di superare gran parte degli ostacoli descritti nel precedente capitolo, in quanto ci si è avvalsi di materiale autentico, omogeneo rispetto a numerose variabili e in quantità sufficienti per essere rappresentativo; lo stesso materiale è stato poi elaborato e reso disponibile in formato elettronico, in modo da poter farne uso a fini di ricerca e didattica attraverso procedure semi-assistite e pertinenti con la linguistica computazionale.

Di seguito sono descritte tutte le fasi della costruzione del corpus che si basa, come primo passo, sulla creazione di un archivio multimediale composto dalle registrazioni audio-video (i discorsi originali), audio (le rese degli interpreti) e dalle trascrizioni del materiale oggetto di studio; il tutto classificato e catalogato in modo tale da avere una gestione funzionale e una reperibilità agile di tutti i tipi di risorse.

La scelta del materiale da includere nello studio è stata guidata da molteplici fattori, quali le fonti disponibili, gli strumenti tecnici più idonei alla raccolta, conservazione ed elaborazione del materiale oggetto di studio e le risorse tecnologiche disponibili al momento dell'attivazione del progetto presso il Dipartimento SITLeC. In seguito a varie valutazioni e grazie al supporto dei tecnici del Dipartimento e della Scuola Superiore di Lingue Moderne per Interpreti e Traduttori (SSLMIT), per raccogliere le prestazioni degli interpreti sono state scelte le sedute plenarie del Parlamento europeo trasmesse dal canale satellitare *Europe by Satellite* (EbS). Tale scelta ha assicurato buona rappresentatività dei dati, essendo possibile reperirne in grandi quantità dalla medesima fonte di provenienza. Inoltre, ha permesso il controllo di numerose variabili, nonché l'omogeneità del contesto di provenienza, delle modalità del servizio di interpretazione simultanea e, caratteristica unica nel suo genere, una disponibilità ampia di materiale in tutte le lingue ufficiali dell'Unione, tra cui italiano, inglese e spagnolo. La fonte prescelta, peraltro già sfruttata per lo

sviluppo di materiali da impiegare nella formazione di interpreti (de Manuel Jerez 2003a) e per la realizzazione di corpora multilingue di lingua scritta (Calzada Pérez & Luz 2006), ha permesso di garantire che per ogni testo (discorso orale) emesso in una delle tre lingue citate fosse sempre disponibile una duplice versione interpretata nelle due lingue restanti, permettendo così un'analisi "a più direzioni" dello stesso testo e fra lingue diverse.

3.1 Impostazione dell'Archivio Multimediale

Fin dalla prima registrazione su videocassetta delle trasmissioni oggetto di studio, è emersa la necessità di impostare un sistema efficace di archiviazione dei dati, in modo che questi fossero poi facilmente reperibili. In vista della loro successiva digitalizzazione e suddivisione in clip corrispondenti ai singoli interventi nelle lingue considerate, era doveroso fare in modo che lo stesso sistema di archiviazione fosse compatibile con la creazione di file digitali in diversi formati (video, audio e testo).

Ai fini di una gestione agevole dell'archivio in costante crescita, è stato approntato un sistema di classificazione che permette di risalire facilmente alla collocazione temporale e linguistica del materiale. La sequenza di informazioni inserite nel nome assegnato a ciascun documento raccolto nell'archivio comprende la data in cui si è svolta la plenaria, le lettere "m" o "p" per segnalare se si tratta di una seduta svolta nel mattino o nel pomeriggio, le sigle "org" oppure "int" per distinguere TP (discorsi originali) e TA (discorsi interpretati), infine le sigle "it" oppure "en" oppure "es" come indicazione linguistica per italiano, inglese e spagnolo rispettivamente. Nell'esempio riportato in basso si fa riferimento a un discorso originale in lingua inglese, pronunciato nel corso della seduta mattutina del 10 febbraio 2004, accompagnato dalle due versioni interpretate in italiano e in spagnolo:

10-02-04-m-org-en
10-02-04-m-int-en-it
10-02-04-m-int-en-es

Oltre alle videocassette e ai relativi file video in formato digitale (§3.2.2), nell'archivio sono raccolte anche tutte le clip editate da tali video, corrispondenti ai singoli interventi pronunciati in italiano, in inglese o in spagnolo con le rispettive versioni interpretate. L'individuazione dei singoli interventi nelle tre lingue di nostro interesse è stata facilitata dalla disponibilità dei processi verbali di ogni seduta plenaria (Marzocchi 2007). Questi documenti, anch'essi

scaricabili da internet e immagazzinati in un'apposita directory dell'archivio multimediale EPIC, hanno la particolarità di segnalare con un codice numerico progressivo l'inizio di ogni sezione di testo. Ad esempio, se il titolo della sessione di apertura di una seduta riporta il codice 001, l'argomento all'ordine del giorno posto come titolo nel verbale in una nuova sezione di testo ha il codice 002. A questo seguono gli interventi dei singoli oratori ai quali viene concessa la facoltà di parola; ogni paragrafo è indicato inizialmente dal cognome di chi interviene, dalla sigla del suo partito politico (o istituzione di afferenza nel caso dei membri della Commissione e del Consiglio) e dal codice numerico appena menzionato, come esemplificato nella Figura 3.1:

Figura 3.1 Estratto del processo verbale della seduta PE del 10/02/2004.

2-002

VORSITZ: GERHARD SCHMID
Vizepräsident

(Die Sitzung wird um 9.00 Uhr eröffnet.)[1]

2-003

Gripe aviar en Asia

Fugleinfluenza i Asien

Vogelgrippe in Asien

Η γρίπη των πουλερικών στην Ασία

Asian bird flu

Grippe aviaire en Asie

Influenza aviaria in Asia

Vogelpest in Azië

Gripe das aves na Ásia

Lintuinfluenssa Aasiassa

Fågelinfluensan i Asien

2-004

Der Präsident. – Nach der Tagesordnung folgt die Erklärung der Kommission zur Vogelgrippe in Asien.

2-005

Byrne, *Commission.* – Mr President, avian influenza has affected poultry populations in nine countries in Asia so far, namely Cambodia, the People's Republic of China, including the special administrative region of Hong Kong, Indonesia, Japan, Laos, Pakistan, South Korea, Thailand and Vietnam.

According to an FAO-WHO-OIE expert panel held on 3 and 4 February in Rome, the current epidemic is still evolving and not yet under control. As of today, the disease has taken 18 human lives in Vietnam and Thailand, but a new virus, fully transmissible to humans and capable of spreading from human to human, has not yet emerged.

Il sistema di codici numerici visibile nel margine sinistro della Figura 3.1 ha facilitato l'individuazione di singoli brani di testo all'interno dei verbali di ogni seduta. Lo stesso codice si è rivelato pertanto estremamente utile al fine di archiviare i dati in tutti i formati da noi considerati (multimediale e testuale), come illustrato nel seguente esempio, tratto dall'intervento 005 del commissario David Byrne sopra riportato:

10-02-04-m-005-org-en.mpeg
10-02-04-m-005-int-en-it.wav
10-02-04-m-005-int-en-es.wav

10-02-04-m-005-org-en.txt
10-02-04-m-005-int-en-it.txt
10-02-04-m-005-int-en-es.txt

Nell'archivio multimediale sono dunque raccolte tutte le clip editate dalle registrazioni delle sedute plenarie, assieme alle registrazioni complete ricavate dalle singole videocassette. Il materiale disponibile ammonta a 103 clip in italiano, 308 clip in inglese e 130 clip in spagnolo. Questi dati corrispondono al numero di clip audio-video dei discorsi originali, che sono sempre corredati da due clip audio con le registrazioni degli interpreti. Pertanto, il dato finale del numero totale di clip raccolte nell'archivio si ottiene moltiplicando il tutto per tre.

Tabella 3.1 Clip raccolte nell'archivio multimediale EPIC.

MATERIALE	ORG-IT	ORG-EN	ORG-ES
clip TP	103	308	130
clip TA(1)	103	308	130
clip TA(2)	103	308	130
TOTALE org+int	**309**	**924**	**390**

L'archivio si compone di un registro impostato in un foglio di lavoro Excel, in cui sono indicati gli argomenti relativi a ogni gruppo di clip con indicazioni sulla durata, sulla collocazione e sul numero di clip disponibili per ogni lingua. Tutto il materiale nelle altre lingue in uso al Parlamento europeo, al momento escluse dalla ricerca, è conservato all'interno dell'archivio per eventuali studi futuri o per possibili impieghi in ambito didattico.[1] Oltre a questo, sono state registrate e archiviate anche varie conferenze stampa che vedono l'uso di diverse lingue. Infine, anche tutti i verbali delle sedute registrate fanno parte dell'archivio e sono raccolti in una directory apposita.

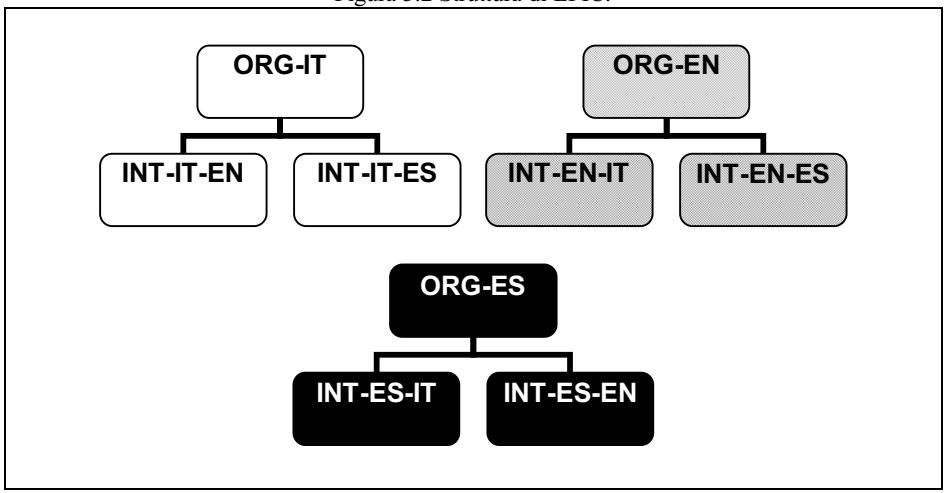
[1] A seguito di alcune richieste da parte di laureandi interessati, sono state editate e fornite le clip dei discorsi disponibili in tedesco e in francese come materiale di studio per le loro tesi di laurea. Inoltre, alcune clip sono state editate in modo da comporre unità di ascolto per un corso di *Sound Perception Training* (Kaunzner 1997).

3.2 Creazione del corpus

3.2.1 Struttura e rappresentatività del corpus

EPIC si compone di nove sottocorpora: tre sottocorpora con i TP e sei sottocorpora con i TA. La struttura globale del corpus è rappresentata nella Figura 3.2 sotto riportata (org = originale, TP; int = interpretazione, TA; IT = italiano; EN = inglese; ES = spagnolo):

Figura 3.2 Struttura di EPIC.



Come si vede chiaramente dall'articolata struttura del corpus, EPIC può essere inquadrato sia come corpus parallelo con TP e TA a confronto, sia come corpus comparabile; ad esempio, i TP italiani possono essere messi a confronto con i TA italiani che sono stati prodotti a partire da TP inglesi o spagnoli. Questa doppia prospettiva di analisi è stata già impiegata in alcune ricerche sulla densità lessicale (Russo et al. 2006; Sandrelli et al. 2010), e su alcuni tipi di disfluenze (Bendazzoli et al. in corso di stampa 2011), a cui si aggiungono due studi in prospettiva parallela sulla coppia di lingue spagnolo-italiano (Russo 2007, 2008).

Al momento, EPIC comprende 119 TP (81 in inglese, 17 in italiano e 21 in spagnolo), ai quali si aggiungono i rispettivi TA per un totale di 357 testi. La dimensione totale del corpus è di circa 177.000 parole, distribuite (in modo

ancora disomogeneo) tra i nove sottocorpora, come rappresentato schematicamente nella Tabella 3.2:

Tabella 3.2 Dimensione di EPIC (prosepttiva parallela).

sottocorpus	numero di testi	numero di parole	% di EPIC
Org-it	17	6.765	4
Int-it-en	17	6.708	4
Int-it-es	17	7.052	4
Org-en	81	42.705	25
Int-en-it	81	35.765	20
Int-en-es	81	38.066	21
Org-es	21	14.406	8
Int-es-en	21	12.995	7
Int-es-it	21	12.833	7
TOTALE	357	177.295	100

Il corpus è attualmente in fase di espansione. Nuove trascrizioni sono state ottenute dal materiale registrato e conservato nell'Archivio Multimediale di EPIC, buona parte del quale non era stato inizialmente inserito nel corpus. Tale materiale è stato di fatto utilizzato da numerosi studenti di interpretazione presso la SSLMIT di Forlì per le loro tesi di laurea. Da questo punto di vista, la creazione di EPIC ha avuto un impatto positivo non solo sulla ricerca, ma anche sulla didattica e sulla formazione di nuovi interpreti (Russo 2010).

3.2.2 Raccolta dei dati

Come segnalato in apertura di questo capitolo, la scelta di raccogliere i dati da una fonte istituzionale quale il Parlamento europeo è stata dettata da molteplici fattori. Oltre a quelli già menzionati, di carattere più organizzativo, va ribadito il vantaggio di poter accedere ai dati da un canale esterno, cioè la televisione

satellitare, nonché di poter operare sulla base del consenso concesso dai responsabili dei servizi audiovisivi del PE.[2]

Sulla base del palinsesto proposto dal canale satellitare EbS, sono state effettuate registrazioni su videocassetta VHS da quattro diverse postazioni televisive con ricevitore satellitare presso il Dipartimento SITLeC e la Facoltà SSLMIT a Forlì. Il canale EbS permette, infatti, di impostare la lingua di trasmissione nel corso delle sedute plenarie del Parlamento europeo, in cui tutte le lingue ufficiali dell'Unione europea sono garantite.[3] Nel nostro studio, le quattro postazioni sono state impostate sui canali di trasmissione audio originale (*floor*), italiano, inglese e spagnolo. Le registrazioni sono state eseguite nei mesi di febbraio, marzo, aprile e luglio dell'anno 2004, per un totale di 22 giornate. Tutto il materiale costituisce già di per sé un archivio audio-video di testi in lingua originale e interpretati simultaneamente, oltre a includere conferenze stampa e materiale informativo sulle attività dell'Unione europea (in totale sono state utilizzate 140 videocassette per 280 ore stimate di materiale utile). In questa fase, inoltre, sono stati attivati preziosi contatti con il personale responsabile degli archivi audiovisivi del Parlamento europeo, grazie al quale è stato possibile ottenere ulteriore materiale e verificare il permesso d'uso dei dati per scopi didattici e di ricerca.

Successivamente, tutto il materiale è stato trasformato in formato digitale e salvato su un disco rigido di un computer dedicato. Il processo di digitalizzazione del materiale audio-video è stato indispensabile per gli innumerevoli vantaggi offerti da questo formato. Va osservato, a questo proposito, che a soli due anni dall'attivazione del progetto di ricerca (nel gennaio del 2004) erano già reperibili sul mercato strumenti validi che consentono di registrare direttamente in formato digitale, ovvero DVD-recorder con hard disk interni in grado di immagazzinare intere giornate di registrazioni già in formato digitale. Si ritiene necessario inserire questo appunto per segnalare quanto il processo di trasformazione del formato delle registrazioni sia stato tanto necessario quanto impegnativo, data la grande quantità di dati raccolti in videocassetta.

La versione originale del materiale è digitalizzata nel formato ".mpeg", consentendo sia la visione del filmato, sia l'ascolto del testo orale. Le versioni interpretate simultaneamente, invece, sono digitalizzate nel formato ".wav", conservando quindi soltanto la traccia audio del materiale.

---

[2] Tutte le registrazioni possono ora essere scaricate dalla pagina web del Parlamento europeo (la biblioteca multimediale del PE contiene le registrazioni a partire da aprile 2006) e il consenso all'uso di questi materiali è concesso a fini non commerciali. Ben diversa è la situazione se si considera la raccolta dei dati in un convegno svolto nel mercato privato (si veda il capitolo successivo, §4.2.2).

[3] Le sedute plenarie del PE prevedono il servizio di interpretazione simultanea in tutte le lingue ufficiali dell'Unione (attualmente ventitré lingue, per un totale di 506 combinazioni linguistiche) conformemente all'Articolo 146 del Regolamento del Parlamento europeo.

La scelta di questi formati, così come per tutta l'impostazione che è stata data al lavoro nella fase iniziale, è frutto di un'analisi attenta di un'ampia gamma di applicazioni software in cui il materiale potrebbe essere esportato senza rischi di incompatibilità, permettendo così futuri percorsi di ricerca paralleli.

Come anticipato nella descrizione del sistema approntato per archiviare i dati registrati, dai filmati interi in formato digitale delle sedute plenarie sono state selezionate le parti di interesse per la ricerca, ovvero tutti i discorsi emessi nelle lingue italiano, inglese e spagnolo e le corrispondenti versioni interpretate nelle tre lingue coinvolte. È stato creato, quindi, un archivio multimediale che raccoglie le clip video dei discorsi in lingua originale e le relative clip audio con le versioni interpretate simultaneamente.

3.2.3 Trascrizione

Sulla base delle considerazioni teorico-pratiche suesposte (§1.3.3), i seguenti tre livelli di trascrizione sono stati considerati nel presente studio: livello linguistico, livello paralinguistico e livello extralinguistico. Data la scarsità di corpora elettronici in questo campo o di grandi quantità di materiale omogeneo già trascritto, nella prima fase della ricerca ci si è concentrati principalmente sul livello verbale/linguistico, ovvero sulle parole emesse dagli oratori e dagli interpreti. Altre caratteristiche, quali ad esempio quelle legate alla prosodia degli oratori e degli interpreti, in riferimento a fenomeni come allungamenti vocalici, cambiamenti nel ritmo d'eloquio e altri, non sono stati considerati. Pur rappresentando un limite, questo è stato necessario in quanto prima di annotare ulteriori fenomeni particolari, si ritiene opportuno studiarli approfonditamente nella letteratura esistente, in modo da procedere sulla base di solide definizioni metodologiche; operazione, questa, difficilmente realizzabile all'inizio di uno studio che ha mirato a preparare una grande quantità di dati e che è stato condotto da un numero limitato di persone.

Inoltre, si ritiene che il processo di annotazione vera e propria di particolari fenomeni debba essere eseguito con l'ausilio di adeguati strumenti informatici, in grado di offrire una rappresentazione più attendibile dal punto di vista scientifico dell'oggetto di studio (ad esempio, per segnalare le pause ci si può avvalere di programmi come *Cooledit*, *Audacity* o *Audiolab* che offrono la rappresentazione visiva delle onde sonore segnalandone eventuali interruzioni e la curva intonativa). Non è sufficiente basarsi sulla sola percezione personale, ovvero l'ascolto, delle registrazioni, in quanto l'orecchio umano e l'attenzione dell'ascoltatore non possono mantenere livelli di percezione costanti nel tempo e potrebbero non cogliere alcuni fenomeni o sopravvalutarne altri. Nelle

trascrizioni di EPIC il metodo basato sulla percezione del trascrittore è stato impiegato limitatamente alle pause, percepite come piene o vuote all'interno del flusso del discorso, che sono state inserite seguendo il secondo principio che ha guidato il processo di trascrizione, ovvero la fruibilità delle trascrizioni da parte di un potenziale lettore e l'agevole realizzazione delle stesse da parte del trascrittore (*user/annotator-friendliness*). In questo senso, le pause così annotate rappresentano un semplice aiuto "visivo" ai potenziali lettori delle trascrizioni, senza per questo assumere alcun valore scientifico per le analisi dei testi in questione. In ogni modo, le trascrizioni così elaborate offrono una base di partenza per futuri studi su caratteristiche del parlato che potranno comunque essere incluse in un secondo momento.

Nel primo livello di trascrizione, ovvero il livello linguistico, ci si è concentrati pertanto sulla realizzazione di trascrizioni ortografiche, seguendo le convenzioni stabilite nella guida in uso presso le istituzioni comunitarie e disponibile in Internet. Si tratta di un documento chiamato "Manuale interistituzionale di convenzioni redazionali" (per l'inglese "Interinstitutional style guide" mentre per lo spagnolo "Libro de estilo interinstitucional", cfr. §Sitografia) in cui sono raccolte le convenzioni ortografiche da seguire nella redazione di documenti per tutte le lingue ufficiali dell'Unione europea.

Le valutazioni a cui si è fatto riferimento sono valide anche e soprattutto per il livello paralinguistico, qui parzialmente considerato. Nello specifico sono stati annotati i fenomeni di troncamento, cioè tutte quelle parole che non sono pronunciate completamente o la cui produzione presenta delle "fratture" interne, e i problemi di disfluenze nella pronuncia, con la conseguente realizzazione di parole inesistenti per quanto comprensibili all'interno del discorso. Nel caso degli errori di pronuncia o dei troncamenti interni, le parole interessate vengono prima "normalizzate", cioè trascritte secondo la grafia corretta, seguite poi dalla forma che riflette la versione orale, secondo una trascrizione letterale. Questo accorgimento è stato necessario per poter annotare automaticamente tutti i *token*[4] dei testi presenti nel corpus (§1.3.4).

Le unità di suddivisione del testo trascritto secondo gli "enunciati" o le "unità di informazione" sono state annotate inserendo una doppia barra all'interno del testo trascritto (//). Tuttavia, oltre a considerare la compiutezza pragmatica (sulla base dell'intonazione e dell'autonomia sintattico-grammaticale), anche il principio della *reader/annotator-friendlyness* ha avuto un peso considerevole nell'espletare questa segmentazione. Le convenzioni stabilite per trascrivere i fenomeni menzionati sono riassunte di seguito nella Tabella 3.3:

[4] L'uso di questo termine è voluto per differenziarlo dal più generico "parola" che può essere composta da più *token* come spiegato al §1.3.4.

Tabella 3.3 Convenzioni di trascrizione in EPIC.

FENOMENI	esempi	CONVENZIONI DI TRASCRIZIONE
Troncamenti finali Troncamenti interni	propo pro posta	propo- proposta </pro_posta/>
Problemi di pronuncia	martalità parlomento	mortalità </martalità/> parlamento </parlomento/>
Pause	(piene / vuote)	ehm ...
Numeri Cifre Date	532 4% 1997	cinquecentotrentadue quattro per cento millenovecentonovantasette
Parole incomprensibili		#
Unità di suddivisione		//

Infine, il livello extralinguistico è stato incluso all'inizio di ogni trascrizione sotto forma di intestazione o *header*, in cui sono fornite informazioni su: situazione di emissione (data), testo (durata in secondi, numero di parole, velocità media calcolata in parole al minuto, argomento trattato, modalità di esposizione) e oratore (nazionalità, lingua, sesso, funzione politica). Tutte le voci qui comprese si sono poi rivelate funzionali alla messa a punto di filtri di ricerca disponibili nell'interfaccia di ricerca automatica *on-line* che consente di interrogare il corpus (§3.2.6). Per ogni campo sono state studiate e selezionate le informazioni da inserire, includendo un ultimo punto per i commenti, dove poter annotare informazioni aggiuntive o di natura tale da non poter essere incluse nei precedenti campi. Nello schema alla Figura 3.3 è riportato un esempio di *header* contenente informazioni su un discorso pronunciato da un europarlamentare italiano. Tale esempio è seguito dalla Tabella 3.4 in cui sono riportati tutti i riferimenti per la compilazione completa dell'*header* con le informazioni pertinenti.

Figura 3.3 Esempio di *header* in EPIC.

```
(date: 10-02-04-m
speech number: 011
language: it
type: org-it

duration: medium
timing: 136

text length: short
number of words: 266

speed: low
words per minute: 117

source text delivery: impromptu

speaker: Fiori, Francesco
gender: M
country: Italy
mother tongue: yes

political function: MEP
political group: PPE-DE

topic: Health
specific topic: Asian bird flu

comments: NA)
```

I vari campi dell'*header* raccolgono diverse informazioni che sono state raggruppate e strutturate nel seguente schema (Tabella 3.4):

Tabella 3.4 *Header* di EPIC: schema di compilazione.

Date	dd-mm-yy-m/p
speech number	
language	it / en / es
type	org-xx / int-xx-xx
duration	short (< 120 secs) medium (121 – 360 secs) long (> 360 secs)
timing	total seconds
text length	short (< 300 words) medium (301 – 1000 words) long (long > 1000)

number of words		
speed	slow medium high	(<100 w/m) (100 - 120 w/m) (> 120 w/m)
words per minute		
source text delivery	impromptu / read / mixed	
speaker	surname, first name	
gender	F / M	
country		
mother tongue	yes / no	
political function	MEP MEP Chairman of the session President of the European Parliament Vice-President of the European Parliament European Commission European Council guest	
political group	Verts/ALE	(Gruppo Verde/Alleanza libera europea)
	PPE-DE	(Gruppo del Partito Popolare Europeo e Democratici Europei)
	PSE	(Gruppo del Partito Socialista Europeo)
	ELDR	(Gruppo del Partito europeo dei liberali, democratici e riformatori)
	GUE/NGL	(Gruppo confederale della Sinistra unitaria europea/Sinistra verde nordica)
	UEN	(Gruppo "Unione per l'Europa delle Nazioni")
	TDI	(Gruppo tecnico dei deputati indipendenti – Gruppo misto)
	EDD	(Gruppo per l'Europa delle democrazie e delle diversità)
	NI	(Non iscritti)
topic	Agriculture & Fisheries Economics & Finance Employment Environment Health Justice Politics Procedure & Formalities Society & Culture Science & Technology Transport	
specific topic	(dicitura ufficiale dal verbale)	

comments	configurazione Consiglio	General Affairs and External Relations Economic and Financial Affairs Cooperation in the fields of Justice and Home Affairs Employment, Social Policy, Health and Consumer Affairs Competitiveness Transport, Telecommunications and Energy Agriculture and Fisheries Environment Education, Youth and Culture
	DG Commissione	Agriculture and Fisheries Administrative Reform Competition Enterprise and Information Society Internal Market Research Development and Humanitarian Aid Enlargement External Relations Trade Health and Consumer Protection Education and Culture Budget Environment Justice and Home Affairs Employment and Social Affairs Regional Policy Economic and Monetary Affairs Relations with the European Parliament, Transport and Energy President of the European Commission
	carica di guest	eg. President of the Republic of Colombia
	accenti	Scottish accent Welsh accent Irish accent Andalusian accent Latin American accent…
	problemi tecnici	breathe in technical problems between 1.16- 1.30
	altro	

I vari parametri utilizzati per classificare la durata, la lunghezza in numeri di parole e la velocità per numero di parole al minuto sono stati stabiliti considerando il particolare contesto di provenienza del materiale. Termini come "lungo", "medio" o "breve" per la lunghezza e la durata, oppure "elevata", "media" o "bassa" per la velocità non hanno valore assoluto e necessitano di un termine di riferimento con il quale poter essere misurati. In questo senso, è stata effettuata una stima delle tendenze risultanti dalle trascrizioni completate per le prime giornate di incontri del mese di febbraio. Pur non fornendo un valore assoluto, queste hanno permesso di delineare un quadro generale del contesto

"Parlamento europeo" in cui una velocità d'eloquio di 130 parole al minuto può essere considerato lo standard a cui gli interpreti si trovano solitamente esposti, con una prevalenza di interventi dalla durata non superiore ai 6 minuti.

Allo stato attuale, il materiale relativo al mese di febbraio è stato totalmente trascritto. In totale si tratta di oltre 20 ore di registrazione, corrispondenti a oltre 177.000 *token*. Il corpus è comunque in continua espansione: infatti, questo è un primo gruppo di testi ai quali si aggiungeranno man mano quelli delle sedute plenarie del Parlamento europeo tenutesi nei mesi successivi, le cui trascrizioni sono ora in fase di revisione per essere poi annotate e inserite nel corpus.[5]

Il processo di trascrizione è stato notevolmente impegnativo, in termini sia di tempo sia di attenzione, e ha richiesto uno sforzo coordinato e attento ai minimi dettagli. Bisogna riconoscere, a questo proposito, che la scelta del materiale oggetto di studio si è rivelata particolarmente vantaggiosa, poiché sono disponibili in Internet tutti i verbali delle sedute del Parlamento, contenenti informazioni sugli oratori e il tema trattato, nonché i titoli specifici degli argomenti e dei documenti all'ordine del giorno. Data la varietà dei temi, questi sono stati raggruppati in macrocategorie stabilite sulla base delle categorie già definite in altre interfacce di ricerca, come quella del portale IATE (*InterActive Terminology for Europe*, già Eurodicautom), o nelle pagine di vari siti informativi sulle attività del Parlamento, della Commissione e dell'Unione europea in generale. I verbali sono stati raccolti per poter essere utilizzati come prima bozza delle trascrizioni degli originali.

Per le interpretazioni, grazie alla nostra formazione da interpreti, ci siamo potuti avvalere della tecnica dello "*shadowing*" (Lambert 1992, Schweda Nicholson 1990), vale a dire l'ascolto delle registrazioni e la simultanea ripetizione a voce alta dello stesso testo, per trascrivere automaticamente i testi utilizzando un software di riconoscimento vocale (sono stati utilizzati due differenti software, *Dragon Naturally Speaking* e *ViaVoice*). Ciò ha consentito un considerevole risparmio di tempo per ottenere una prima bozza di trascrizione dei testi interpretati. Le trascrizioni, infine, sono sempre state sottoposte a controllo incrociato e riviste minuziosamente per ridurre al minimo le imprecisioni. Alcune studentesse laureande presso la SSLMIT di Forlì hanno contribuito alla realizzazione delle trascrizioni utilizzando parte del materiale per le loro tesi di laurea.[6] Anche nel caso delle trascrizioni, il formato scelto per

[5] Da un calcolo approssimativo, il corpus dovrebbe aumentare di oltre 278.000 parole, arrivando quindi a più di 450.000 parole totali, con un incremento diverso a seconda della lingua considerata. Solo per i sottocorpora di TP, è prevista la seguente espansione: org-it +19.800 parole; org-en +79.400 parole; org-es +3.800 parole.

[6] Ad oggi sono state realizzate oltre dieci tesi di laurea utilizzando materiali dell'archivio multimediale EPIC.

questo tipo di documenti vuole essere il più flessibile possibile perché possano essere utilizzati in diverse applicazioni (formato ".txt").

3.2.4 Codifica e annotazione

Le trascrizioni salvate in formato testo e raccolte nell'archivio multimediale devono essere elaborate in un formato compatibile con i programmi che effettuano annotazioni automatiche e l'indicizzazione, consentendone l'esplorazione automatica. Nel nostro caso, i programmi utilizzati per l'annotazione della parte del discorso (*POS-tagging*) e del lemma sono Treetagger (Baroni et al. 2004, Schmid 1994) per i testi italiani e inglesi e Freeling (Carreras et al. 2004) per i testi spagnoli. Come spiegato nel primo capitolo (§1.3.4), si tratta di applicazioni informatiche che assegnano automaticamente "etichette" con le categorie grammaticali alle parti del discorso sulla base di regole grammaticali e statistiche. Le varie etichette costituiscono un repertorio di sigle apparentemente complesse, ma funzionali all'estrazione automatica di informazioni linguistiche dal corpus.

I testi così annotati possono essere analizzati automaticamente, nel nostro caso con i programmi offerti dall'*IMS Corpus Work Bench* (CWB), sviluppato presso l'*Institute for Natural Language Processing* dell'Università di Stoccarda (Christ 1994). Tali programmi consentono di interrogare il corpus attraverso *queries* (richieste di informazione) formulate secondo le speciali regole sintattiche del linguaggio *Corpus Query Processor* (CQP). In questo modo è possibile estrarre automaticamente informazioni linguistiche dal corpus, aumentandone in maniera sostanziale le possibilità esplorative.

Figura 3.4 Esempio di trascrizione "taggata".

<speech date="10-02-04-m" id="005" lang="it" type="int-en-it" duration="long" timing="392" textlength="medium" length="899" speed="medium" wordsperminute="138" delivery="read" speaker="interpreter" gender="M" country="NA" mothertongue="yes" function="NA" politicalgroup="NA" gentopic="Health" sptopic="Asian bird flu" comments="NA">
...

//	//	//	SENT
è	è	essere	VER:pres
una	un	una	DET:indef
malattia	malattia	malattia	NOM
devastante	devastante	devastante	ADJ
con	con	con	PRE
un'	un'	un'	DET:indef
alta	alta	alto	ADJ
mortalità	/martalità/	mortalità	NOM
fino	fino	fino	CON
al	al	al	PRE:det
novantacinque	novantacinque	UNKNOWN	NOM
per	per	per	PRE
cento	cento	cento	ADJ
e	e	e	CON
in	in	in	PRE
ventiquattro	ventiquattro	ventiquattro	ADJ
ore	ore	ora	NOM
//	//	//	SENT

Nell'esempio sopra riportato alla Figura 3.4 è possibile vedere nella prima colonna a sinistra il testo della trascrizione tokenizzato, ovvero con ogni singola parola incolonnata in un'unica stringa verticale; nella colonna successiva si vedono le parole nel formato "trascrizione" in cui eventuali annotazioni di fenomeni di errori di pronuncia o troncamenti interni possono essere visualizzati (racchiusi tra due barre, come nel caso di /martalità/); la terza colonna corrisponde al lemma di ciascun termine; nella quarta colonna sono elencati i POS-tag, ovvero le etichette con le sigle che indicano le parti del discorso a cui appartengono le parole (ADJ sta per "aggettivo", NOM sta per "nome" ecc). Ogni *tagger* ha le proprie liste di abbreviazioni che possono essere poi combinate in sequenze particolari da ricercare all'interno del corpus.

Si ritiene opportuno segnalare che i *tagger* utilizzati non sempre assegnano l'etichetta corretta ai *token* presenti nel corpus. Oltre a problemi derivanti dalla presenza di termini tecnici o nomi stranieri non inclusi nel vocabolario del *tagger*, sono le caratteristiche stesse del testo "trascritto" a far vacillare in alcuni casi il corretto funzionamento del *tagger*. Si tratta, in effetti, di testi che riflettono da vicino l'oralità, mantenendo intatte tutte quelle caratteristiche e quei fenomeni solitamente assenti dai testi di natura propriamente scritta. Alcuni esempi possono essere le false partenze o le ripetizioni, le quali rendono meno lineare la struttura sintattica del testo. I *tagger* sono stati messi a punto pensando a testi di natura scritta che, per quanto complessi, rispettano determinate norme. Quindi, di fronte a fenomeni più legati

all'oralità, i *tagger* potrebbero "interpretare" male la sequenza delle parti del discorso coinvolte e questo potrebbe influire negativamente su tutta la stringa di elementi successivi. In particolare, nelle trascrizioni di EPIC la segnalazione delle pause con le convenzioni stabilite ("…" per le pause vuote e "ehm" per le pause piene) non è riconosciuta dal *tagger* che assegna l'etichetta statisticamente più plausibile ma scorretta, indebolendo la sequenza di assegnazione correttamente iniziata. A fronte di queste incongruenze, nei primi studi condotti sul materiale incluso nel corpus è stato necessario procedere alla verifica manuale dell'assegnazione dell'etichetta corretta, caso per caso, nelle liste di frequenza. Tuttavia, in termini generali, il livello di correttezza nell'assegnazione delle etichette si è dimostrato più che soddisfacente (Sandrelli & Bendazzoli 2006), ottenendo percentuali superiori al 90% in tutti i sottocorpora.

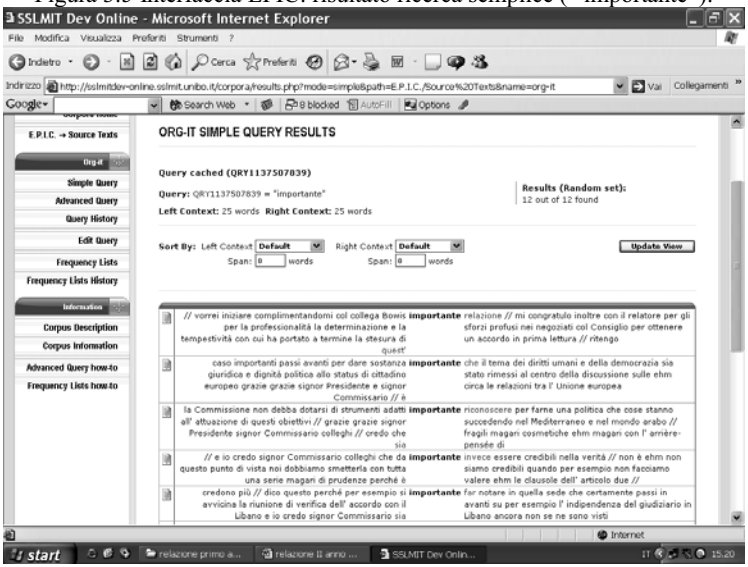
3.2.5 Allineamento

Questa fase rappresenta uno dei prossimi sviluppi del corpus (Russo et al. 2010). A questo proposito, sono stati valutati due programmi, quali Transana e SpeechIndexer, e i relativi sistemi di allineamento (§1.3.5). Tuttavia, l'articolata struttura del corpus (con ben nove sottocorpora e la possibilità di un triplice allineamento) comporta sfide ancora maggiori rispetto all'allineamento di un corpus semplicemente bilingue come DIRSI-C (§4.2.5), dovendo coordinare tre diversi testi trascritti e due diverse tipologie di file multimediali (video per i TP e audio per i TA). Dall'altra parte, la fase di allineamento testo-suono e TP-TA (sulla base del contenuto) è già stata percorsa per DIRSI-C e l'esperienza acquisita può sicuramente fungere da base per ulteriori sviluppi, applicabili al corpus EPIC.

3.2.6 Accessibilità al corpus

Oltre a poter essere analizzato con gli strumenti sopra menzionati ed esplorando il suo archivio multimediale, EPIC è accessibile alla pagina web <http://sslmitdev-online.sslmit.unibo.it/>. Si tratta di un portale in cui sono presenti numerose risorse per interpreti e traduttori professionisti, studenti di interpretazione e traduzione, linguisti, terminologi e ricercatori. Per quanto riguarda EPIC, sono i testi trascritti con le relative annotazioni a poter essere interrogati. Al fine di svolgere analisi automatiche, grazie al supporto del personale tecnico della SSLMIT è stata approntata un'interfaccia di ricerca *ad hoc*. L'interfaccia contiene varie pagine che descrivono ampiamente la struttura e le caratteristiche di EPIC, nonché una maschera di ricerca che consente di accedere al corpus attraverso delle *queries* che permettono di estrarre informazioni dalle trascrizioni caricate attualmente nel corpus. È possibile perlustrare il corpus attraverso ricerche semplici o avanzate. Nel caso di una ricerca semplice, dopo aver impostato i filtri a disposizione che riprendono i parametri descritti in merito all'*header* presente in ogni trascrizione, basta digitare la parola o la stringa di parole che si desidera ricercare nell'apposita cella. Il risultato mostrerà l'elemento ricercato in modalità KWIC (*key word in context*), offrendo inoltre la possibilità di risalire al testo completo della trascrizione in cui è stato riscontrato quel termine particolare (si veda la Figura 3.5 per un esempio di visualizzazione):

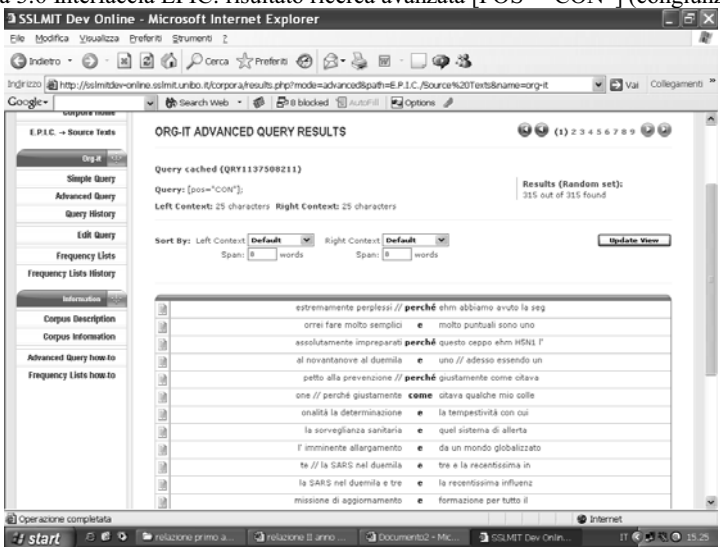
Figura 3.5 Interfaccia EPIC: risultato ricerca semplice (“importante”).



Al momento, solamente dai computer collegati alla rete del dipartimento SITLeC è anche possibile scaricare la clip del discorso originale o interpretato corrispondente, facilmente recuperabile grazie al sistema di archiviazione basato sulla catalogazione di tutto il materiale con il codice precedentemente descritto (§3.1).

Nel caso di una ricerca avanzata, invece, è necessario conoscere la speciale sintassi CQP (*Corpus Query Processor*) e digitare la formula corrispondente alla particolare ricerca che si desidera effettuare con gli strumenti raggruppati in CWB (*Corpus Work Bench*). Questa consente di richiamare occorrenze di complesse stringhe di elementi, corrispondenti al tipo di annotazione applicata al corpus. Nel nostro caso, è possibile per esempio richiamare tutte le parole che presentano problemi di pronuncia, inserite fra due barre, e verificare a cosa sono dovute determinate difficoltà di produzione nel testo orale. Oppure, è possibile studiare determinati prefissi, suffissi, precise costruzioni sintattiche (Bowker 2002, Laviosa 2002), combinando tra loro, ad esempio, parole, lemmi e POS-tag. Nell’esempio sotto riportato (Figura 3.6) è mostrata la schermata con i risultati dell’estrazione di tutte le congiunzioni presenti nel sottocorpus di TP italiani. Al fine di estrarre tali occorrenze, è bastato formulare la richiesta inserendo l’etichetta “CON”, assegnata a tutti i tipi di congiunzioni, quale che sia la loro forma grafica. In assenza del POS-tagging, sarebbe stato necessario svolgere tante ricerche quanti sono i diversi tipi di congiunzioni presenti nel corpus:

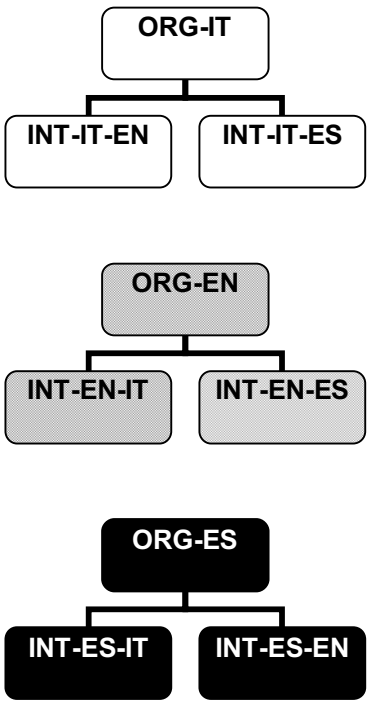
Figura 3.6 Interfaccia EPIC: risultato ricerca avanzata [POS= "CON"] (congiunzione)



Attualmente, uno dei principali limiti dell'interfaccia di ricerca EPIC risiede nell'impossibilità di esportare i risultati ottenuti da una query e di poterli gestire in un documento autonomo (per esempio un foglio di lavoro Excel). Lo stesso limite non si pone se la ricerca è effettuata in ambiente Unix (usando un client SSH come Putty), sfruttando la riga di comando e la suite di programmi della CWB. Ad ogni modo, l'attenzione posta nella scelta dei formati e nell'impostazione del corpus stesso apre ad ulteriori possibilità di sfruttamento di questa risorsa. Un esempio emblematico è dato dalla collaborazione instaurata con un gruppo di ricercatori della *Bar-Ilan University*, i quali hanno impostato una nuova interfaccia con funzionalità alternative (Shlesinger & Ordan 2010). Questo scambio ha confermato l'ottimo grado di esportabilità dei materiali raccolti in EPIC, che resta una risorsa disponibile gratuitamente a tutti coloro che sono impegnati in attività di ricerca e didattica nell'ambito dell'interpretazione e non solo (§5).

3.3 Descrizione di EPIC

EPIC si compone di 9 sottocorpora, di cui 3 con le trascrizioni di discorsi originali in italiano, inglese e spagnolo e 6 con le trascrizioni delle rese degli interpreti, in una struttura che copre tutte le direzioni possibili fra le tre lingue coinvolte. Con la sigla ORG si fa riferimento ai discorsi originali, mentre la sigla INT si riferisce alle rese degli interpreti (per comodità riportiamo nuovamente la Figura 3.2 con una disposizione verticale dei tre gruppi di sottocorpora):



Nel biennio di ricerca in cui il corpus è stato creato, ci si è concentrati su un primo gruppo di trascrizioni, corrispondenti ad alcune parti di cinque sedute plenarie tenute nel mese di febbraio 2004, per un totale di 119 interventi a quali si aggiungono le 238 versioni interpretate. Il numero totale di *token* nel corpus così composto è di oltre 175.000 unità distribuite nei vari sottocorpora (Tabella 3.2). Le dimensioni di ciascun sottocorpus sono in ogni modo destinate ad aumentare considerevolmente con l'aggiunta futura di tutto il materiale raccolto.

3.3.1 Sottocorpora EPIC di discorsi originali

I tre sottocorpora di discorsi originali presentano caratteristiche molto diverse tra loro, prima fra tutte la dimensione. Questa si riflette inevitabilmente anche sui sottocorpora che raccolgono le rese degli interpreti. Le giornate in cui si sono svolte le sedute plenarie del Parlamento europeo incluse nel materiale raccolto per il mese di febbraio 2004 sono elencate nella Tabella 3.5:

Tabella 3.5 Elenco TP disponibili in EPIC per lingua e per seduta.

Data	Numero TP per lingua		
	ORG-IT	ORG-EN	ORG-ES
10-02-04-m	2	17	2
11-02-04-m	4	13	6
12-02-04-m	2	10	3
12-02-04-p	0	10	0
25-02-04-p	7	24	10
26-02-04-m	2	7	0
TOTALE	17	81	21

Come emerge chiaramente dai dati riportati nella Tabella 3.5 e prima ancora nella Tabella 3.2, il sottocorpus di discorsi prodotti in lingua inglese è di dimensioni notevolmente superiori a quelli nelle altre due lingue. Questo dato è spiegabile considerando che la lingua inglese è parlata dai rappresentanti di ben due paesi, Regno Unito e Irlanda, oltre che da altri oratori stranieri che la usano come *lingua franca*; inoltre, la Presidenza di turno dell'Unione nel corso del primo semestre del 2004 spettava all'Irlanda, quindi gli interventi dei rappresentanti del Consiglio erano sempre in lingua inglese (ad eccezione di qualche caso in gaelico). Tuttavia, è doveroso precisare che la distribuzione quantitativa degli oratori non riflette pienamente la situazione del Parlamento, bensì quanto registrato dalle trasmissioni del canale satellitare *EbS* che ricopre solo una parte delle sedute parlamentari.

3.3.1.1 Sottocorpus ORG-IT

Il sottocorpus di discorsi originali in lingua italiana è il più piccolo per dimensione rispetto agli altri due sottocorpora di discorsi originali. Nel mese di febbraio sono stati raccolti 17 interventi per una durata complessiva di circa 50

minuti. Tutti gli interventi sono pronunciati da parlamentari, di cui 14 uomini (soltanto in un caso si tratta dello stesso oratore) e 3 donne, appartenenti ai seguenti partiti politici:

Tabella 3.6 Numero di TP in ORG-IT per gruppo politico.

PARTITO	**N° di interventi**
PPE-DE	6
NI	4
PSE	3
ELDR	2
UEN	2

Alcuni interventi da parte di rappresentanti della Commissione sono sì presenti, ma in altri mesi per ora non inclusi nel corpus (solo presenti nell'archivio multimediale). Ad esempio, nel mese di luglio 2004, in occasione dell'incontro in cui fu celebrato il passaggio della Presidenza di turno dell'UE dall'Irlanda all'Olanda, sono stati registrati numerosi interventi dell'allora Presidente della Commissione Romano Prodi.

Le modalità di emissione del testo di partenza sono state suddivise in tre tipologie: esposizione spontanea o a braccio (*impromptu*), lettura (*read*), oppure "mista" fra le due modalità (*mixed*). In questo gruppo di testi prevale la modalità lettura che assieme a quella mista è quasi il doppio dei casi di esposizione spontanea. Questo dato è in linea con il contesto comunicativo del Parlamento in cui ogni oratore ha un tempo di parola prestabilito che non può essere assolutamente prolungato (in caso di superamento del limite previsto viene fatto un richiamo e viene a breve tolta la parola).

Tabella 3.7 Numero di TP in ORG-IT per modalità di esposizione.

DELIVERY	**N° di interventi**
impromptu	6
read	8
mixed	3

I temi trattati sono vari e in questo caso sono stati raggruppati nelle macrocategorie previste dall'*header* e riassunte nella seguente tabella:

Tabella 3.8 Numero di TP in ORG-IT per area tematica.

TOPIC	N° di interventi
Politics	5
Economics & Finance	4
Justice	4
Health	2
Agriculture & Fisheries	1
Transport	1

Considerando ora gli attributi strettamente legati al testo di partenza vero e proprio, ovvero i discorsi che sono stati poi interpretati per essere fruiti anche dagli altri parlamentari, vengono di seguito riportate le caratteristiche registrate.

La durata in termini di tempo dei discorsi riflette da vicino il ruolo istituzionale degli oratori registrati per questo sottocorpus. Si tratta infatti di europarlamentari, i quali, a differenza dei Commissari o dei rappresentanti del Consiglio, hanno in genere un tempo di parola limitato a qualche minuto per intervenire:

Tabella 3.9 Numero di TP in ORG-IT per durata.

DURATION	N° di interventi
long > 360 sec	0
medium 121-360 sec	13
short < 120 sec	4

La lunghezza degli interventi in termini di parole è in linea con il tempo a disposizione degli oratori:

Tabella 3.10 Numero di TP in ORG-IT per lunghezza (numero di parole).

TEXT LENGTH	N° di interventi
long > 1000	0
medium 301-1000	10
short < 300	7

La velocità media è calcolata in termini di parole al minuto, conseguentemente questo parametro è una funzione dei due precedenti parametri:

Tabella 3.11 Numero di TP in ORG-IT per velocità (parole al minuto).

SPEED	N° di interventi
high > 160w/m	0
medium 130-160 w/m	6
low < 130	11

Il quadro generale delle caratteristiche dei discorsi in italiano vede la presenza di interventi di breve o media durata e lunghezza (sempre in riferimento alle prassi comunicative nel contesto di provenienza dei dati), pronunciati da eurodeputati a una velocità medio-bassa rispetto ai normali ritmi che caratterizzano le sedute plenarie del Parlamento europeo. Il dato sulla velocità di eloquio (con un valore medio di circa 130 parole al minuto) sembrerebbe contrastare con la comune percezione dei relatori italiani come di persone che parlano a una velocità estremamente elevata.

3.3.1.2 Sottocorpus ORG-EN

Il sottocorpus di discorsi originali in lingua inglese è il maggiore per dimensione. Comprende 81 discorsi da parte di oratori provenienti dal Regno Unito, dall'Irlanda e da tre paesi non anglofoni, quali Portogallo, Olanda e Danimarca. La durata totale di tutti questi interventi ammonta a quasi cinque ore.

Tabella 3.12 Numero di TP in ORG-EN per provenienza dell'oratore.

PAESE PROVENIENZA	**N° di interventi**
United Kingdom	43
Ireland	35
Portugal	1
The Netherlands	1
Denmark	1

Si tratta in totale di 65 uomini e 16 donne, con numerosi casi in cui lo stesso oratore prende più volte la parola, soprattutto nel caso del Presidente Patrick Cox (13 interventi) e dei Commissari Christopher Patten (11 interventi) e David Byrne (6 interventi). In dettaglio, gli oratori ricoprono le cariche riportate nella seguente tabella:

Tabella 3.13 Numero di TP in ORG-EN per ruolo istituzionale dell'oratore.

RUOLO ISTITUZIONALE	**N° di interventi**
MEP	42
President of EP	13
Vice-president of EP	1
Commission	18
Council	7

Tranne che per i rappresentanti della Commissione e del Consiglio, l'appartenenza ai vari partiti politici degli oratori è riassunta nella seguente tabella:

Tabella 3.14 Numero di TP in ORG-EN per gruppo politico.

PARTITO	**N° di interventi**
PPE-DE	16
PSE	9
Verts/ALE	5
UEN	4
EDD	1

Fra gli oratori presenti in questo sottocorpus, la modalità di esposizione privilegiata è stata come per gli oratori italiani la lettura. Il numero di esposizioni "a braccio" è comunque significativo non solo per il numero maggiore di interventi qui raccolti, ma anche perché fra questi sono comprese le risposte che i rappresentanti della Commissione o del Consiglio sono chiamati a fornire, durante il turno di replica alla fine di ogni dibattito, ai parlamentari intervenuti precedentemente:

Tabella 3.15 Numero di TP in ORG-EN per modalità di esposizione.

DELIVERY	N° di interventi
impromptu	24
mixed	14
read	43

Fra gli argomenti trattati si nota, in questo caso, anche la presenza della voce "*procedures and formalities*" con numerosi interventi. Questi corrispondono alle formule in uso presso il Parlamento europeo per trasmettere comunicazioni ufficiali a tutti i parlamentari, oppure per aprire o chiudere un dibattito indicandone l'ordine del giorno e quant'altro:

Tabella 3.16 Numero di TP in ORG-EN per area tematica.

TOPIC	N° di interventi
Politics	31
Health	15
Procedure & Formalities	13
Justice	12
Economics & Finance	6
Agriculture & Fisheries	1
Employment	1
Environment	1
Transport	1

Passando alle caratteristiche dei discorsi, queste sono riassunte nella seguente serie di tabelle.

A differenza del precedente sottocorpus italiano, i discorsi in lingua inglese comprendono soprattutto interventi di media e lunga durata. Come già

riscontrato nella descrizione dedicata agli oratori presenti in questo sottocorpus, gli interventi più lunghi sono da attribuire ai Commissari o a rappresentanti del Consiglio che hanno la possibilità di esporre in maniera esaustiva il lavoro svolto sui rispettivi temi di competenza, nonché di rispondere a fine dibattito a tutte le persone che vi hanno partecipato:

Tabella 3.17 Numero di TP in ORG-EN per durata.

DURATION	N° di interventi
long > 360 sec	13
medium 121-360 sec	40
short < 120 sec	28

Il parametro relativo alla lunghezza degli interventi in termini di parole è in linea con il precedente parametro salvo poche eccezioni:

Tabella 3.18 Numero di TP in ORG-EN per lunghezza (numero di parole).

TEXT LENGTH	N° di interventi
long > 1000	10
medium 301-1000	44
short < 300	27

Ben diverso rispetto al caso dell'italiano è il parametro sulla velocità di eloquio in termini di parole al minuto. Prevalgono qui gli interventi pronunciati a una velocità elevata o media, a fronte di un ristretto numero di discorsi caratterizzati da una velocità d'eloquio al di sotto delle 130 parole al minuto. Impressionante è il numero di interventi pronunciati a una velocità superiore alle 160 parole al minuto, certamente non ideale per gli interpreti che tra l'altro devono tenere sotto controllo i passaggi di parola per sintonizzarsi sul canale appropriato al momento opportuno. In generale, la velocità media degli interventi in inglese è di 156,5 parole al minuto:

Tabella 3.19 Numero di TP in ORG-EN per velocità (parole al minuto).

SPEED	N° di interventi
high > 160w/m	34
medium 130-160 w/m	36
low < 130	11

3.3.1.3 Sottocorpus ORG-ES

Il sottocorpus di discorsi in lingua spagnola comprende 21 discorsi, per una durata di quasi due ore, pronunciati in 14 casi da oratori uomini e in 7 da donne. Gli unici casi di interventi da parte della stessa persona si hanno con il Vicepresidente del Parlamento Joan Colom i Naval e la Vicepresidente della Commissione, nonché Commissario Loyola De Palacio Vallelersundi, entrambi con 3 interventi ciascuno (tutte le cariche sono ovviamente riferite al 2004). Inoltre, è presente un caso di un oratore non proveniente dalla Spagna, bensì dalla Colombia; si tratta del Presidente Uribe, ospite al Parlamento nella giornata del 10 febbraio 2004. La composizione specifica degli oratori per questo sottocorpus è riassunta nella Tabella 3.20:

Tabella 3.20 Numero di TP in ORG-ES per ruolo istituzionale dell'oratore.

RUOLO ISTITUZIONALE	N° di interventi
MEP	11
Vice-president of EP	3
Commission	6
guest	1

L'appartenenza politica degli oratori, tranne che per gli esponenti della Commissione e l'ospite, è riportata nella tabella a seguire:

Tabella 3.21 Numero di TP in ORG-ES per gruppo politico.

PARTITO	N° di interventi
PSE	7
PPE-DE	5
ELDR	1
Verts/ALE	1

Le modalità di esposizione dei vari interventi in lingua spagnola vedono la prevalenza sempre della lettura che, aggiunta alla modalità mista, caratterizza gran parte degli interventi.

Tabella 3.22 Numero di TP in ORG-ES per modalità di esposizione.

DELIVERY	N° di interventi
impromptu	5
mixed	7
read	9

Passando ora ai temi discussi negli interventi raccolti, anche qui compare la voce "*procedures and formalities*" dovuta probabilmente alla partecipazione del Vicepresidente spagnolo (l'onorevole Colom i Naval):

Tabella 3.23 Numero di TP in ORG-ES per area tematica.

TOPIC	N° di interventi
Politics	9
Justice	4
Economics & Finance	3
Procedure & Formalities	3
Agriculture & Fisheries	1
Transport	1

I discorsi in spagnolo presentano le caratteristiche raggruppate nelle seguenti tabelle. La durata in termini di tempo degli interventi in spagnolo vede una distribuzione abbastanza omogenea fra interventi lunghi e brevi, con una quota maggiore di interventi di media durata. Vale la pena precisare che tra gli interventi lunghi è compresa l'allocuzione del Presidente Uribe, con la durata di tempo maggiore (poco più di 25 minuti) e il più alto numero di parole (3.189) tra tutti i TP inclusi in EPIC:

Tabella 3.24 Numero di TP in ORG-ES per durata.

DURATION	N° di interventi
long > 360 sec	4
medium 121-360 sec	12
short < 120 sec	5

Il parametro relativo alla lunghezza degli interventi in termini di parole riflette da vicino la situazione inerente al precedente parametro:

Tabella 3.25 Numero di TP in ORG-ES per lunghezza (numero di parole).

TEXT LENGTH	N° di interventi
long > 1000	4
medium 301-1000	10
short < 300	7

Analogamente al dato registrato per il sottocorpus inglese, anche nei discorsi in lingua spagnola si riscontra una prevalenza di velocità elevata o media, con una minima quantità di interventi pronunciati a velocità bassa. Il valore medio di questi interventi è di 152 parole al minuto, ben oltre la soglia tipicamente ritenuta "agevole" per lo svolgimento dell'interpretazione simultanea (100-120 parole al minuto):

Tabella 3.26 Numero di TP in ORG-ES per velocità (parole al minuto).

SPEED	N° di interventi
high > 160w/m	10
medium 130-160 w/m	7
low < 130	4

3.3.2 Sottocorpora EPIC di discorsi interpretati

Tutti i discorsi originali inclusi in EPIC sono stati interpretati in tutte le lingue ufficiali dell'Unione. Nel nostro caso, l'attenzione è stata posta sulle lingue italiano, inglese e spagnolo, ovvero su coppie di lingue affini (italiano e spagnolo) e non affini (inglese e italiano/spagnolo). Gli interpreti che lavorano al Parlamento si trovano a gestire discorsi in genere estremamente veloci e densi. Nonostante il fatto che, da quanto descritto al punto precedente, risulti che gli oratori italiani inclusi nel corpus mantengano una velocità d'eloquio relativamente bassa, la tendenza degli oratori è quella di dire il più possibile nel tempo a loro concesso, consapevoli di non avere possibilità di ricevere generose estensioni della facoltà di parola. Bisogna infatti ricordare che stando a quanto riportato in letteratura, e nell'opinione dei professionisti in generale, la velocità

d'emissione del TP è considerata agevole ai fini dell'IS se rientra tra le 100-120 parole al minuto (Seleskovitch 1978, Shlesinger 2003). Pertanto, la velocità di emissione dei discorsi al Parlamento europeo si caratterizza, assieme ad altre numerose variabili, per condizioni a dir poco "estreme" che sembrano però essere la norma per gli interpreti (Marzocchi e Zucchetto 1997).

Le caratteristiche dei sottocorpora di discorsi interpretati sono presentate di seguito considerando le due diverse direzioni d'arrivo assieme, a partire dalla stessa lingua fonte. Si consideri che si tratta di interpreti diversi, ma è possibile riconoscere la voce dello stesso interprete al lavoro per vari interventi nel corso della stessa giornata o anche in giorni successivi. Quindi, il dato quantitativo corrispondente al numero di discorsi originali non deve far pensare alla presenza di altrettanti interpreti. È bene ricordare che nelle riunioni al Parlamento sono normalmente presenti tre interpreti per cabina (talvolta in plenaria ce ne possono essere anche quattro per coprire una gamma più ampia di combinazioni), operativi in turni che non dovrebbero eccedere le sette ore, spezzati dalla pausa pranzo (Monti, comunicazione personale). Di norma, gli interpreti svolgono un turno di due ore al mattino e un altro nel pomeriggio. Nella giornata di lunedì è previsto soltanto un turno pomeridiano di tre ore e mezza. Ogni squadra di interpreti in genere lavora per un solo giorno a tornata (Marzocchi, comunicazione personale). La stessa precisazione sul numero effettivo di interpreti vale per tutti gli altri dati forniti in merito ai sottocorpora contenenti i discorsi d'arrivo.

3.3.2.1 Sottocorpora INT-IT-EN e INT-IT-ES

I TP registrati in lingua italiana sono interpretati dagli interpreti delle cabine di inglese e spagnolo e le voci percepite dall'ascoltatore indicano la seguente distribuzione di genere:

Tabella 3.27 Numero di interventi interpretati da interpreti uomini o donne dal sottocorpus ORG-IT.

GENDER	INT-IT-EN	INT-IT-ES
M	8	7
F	9	10

La durata in termini di tempo è stata mantenuta uguale al tempo d'esposizione del discorso originale. Pur essendo vero che l'interprete simultaneo non parla esattamente sopra l'oratore, in quanto si registra sempre uno scarto di tempo, il *décalage*, tra l'emissione del discorso originale e l'emissione del discorso tradotto da parte dell'interprete, nel contesto del Parlamento europeo gli

interpreti sono comunque abituati a completare la propria resa assieme o il più vicino possibile alla fine del discorso dell'oratore. Questo è dovuto al fatto che i tempi di parola e i passaggi di turno sono talmente rapidi che anche gli interpreti devono fare il possibile per non "rubare" secondi preziosi ai colleghi di un'altra cabina che devono iniziare subito ad operare non appena il nuovo oratore inizia il discorso.

La lunghezza dei testi in termini di parole, invece, subisce ovviamente dei cambiamenti, e per la natura diversa delle lingue (basta pensare alla gestione dei numeri che in italiano si esprimono con una sola parola nel caso delle centinaia, mentre in inglese e in spagnolo vengono composti da più parole separate tra loro), e per il processo traduttivo, quale possibile indicazione di strategie o problemi.

Di fronte agli stessi discorsi e oratori, in inglese è stato registrato un lieve incremento in un paio di casi, fatto alquanto singolare considerata la tendenza solitamente opposta dell'italiano a essere più verboso e meno sintetico dell'inglese:

Tabella 3.28 Numero di interventi in INT-IT-EN e INT-IT-ES per lunghezza (numero di parole).

TEXT LENGTH	ORG-IT	INT-IT-EN	INT-IT-ES
long > 1000	0	0	0
medium 301-1000	10	12	10
short < 300	7	5	7

Il parametro velocità registra, invece, cambiamenti più marcati, con un generale aumento per la cabina inglese, mentre sembra esserci stato più che altro una diminuzione di velocità (tranne che per un caso) per la cabina spagnola:

Tabella 3.29 Numero di interventi in INT-IT-EN e INT-IT-ES per velocità (parole al minuto).

SPEED	ORG-IT	INT-IT-EN	INT-IT-ES
high > 160 w/m	0	1	1
medium 130-160 w/m	6	8	3
low < 130	11	8	13

È interessante notare che a differenza degli altri sottocorpora, la velocità media di entrambe le cabine qui considerate è superiore alla velocità media dei rispettivi TP (130 parole al minuto): 132,2 parole per la cabina inglese e 136 parole al minuto per la cabina spagnola.

Un altro dato in controtendenza riguarda la lunghezza dei testi, cioè il numero di parole prodotte. Innanzitutto si noti che il sottocorpus contenente i TA inglesi è il più piccolo per dimensione se si considera il numero di parole in tutti i sottocorpora che compongono EPIC. In particolare, gli interpreti inglesi hanno subito una leggera flessione (-57 parole) rispetto ai TP italiani, mentre gli interpreti spagnoli hanno addirittura prodotto un numero di parole superiore (+287 parole). Questo potrebbe forse essere spiegato dalla vicinanza tra le due lingue romanze, rispetto alle maggiori differenze presenti nella direzione di lavoro italiano>inglese, ovvero da lingua romanza a lingua germanica.

3.3.2.2 Sottocorpora INT-EN-IT e INT-EN-ES

Il gruppo di discorsi in lingua inglese è interpretato dagli interpreti delle cabine di italiano e spagnolo secondo una distribuzione di genere che vede la prevalenza di voci femminili in entrambi i casi, specialmente per la cabina italiana:

Tabella 3.30 Numero di discorsi interpretati da interpreti uomini o donne dal sottocorpus ORG-EN.

GENDER	INT-EN-IT	INT-EN-ES
M	13	23
F	68	58

Il parametro relativo alla lunghezza dei TA in termini di parole non presenta variazioni rilevanti tra le due cabine. Soltanto rispetto al dato registrato per i discorsi originali si riscontra una diminuzione del gruppo di testi di media lunghezza a favore dei testi di lunghezza inferiore:

Tabella 3.31 Numero di interventi in INT-EN-IT e INT-EN-ES per lunghezza (numero di parole).

TEXT LENGTH	ORG-EN	INT-EN-IT	INT-EN-ES
long > 1000	10	9	10
medium 301-1000	44	34	33
short < 300	27	38	38

Nel parametro sulla velocità misurata in parole al minuto emergono dati molto diversi tra loro, con una sostanziale diminuzione della velocità nella cabina

italiana e una diminuzione un po' meno marcata nella cabina spagnola, pur essendo due lingue affini provenienti dalla stessa lingua di partenza:

Tabella 3.32 Numero di interventi in INT-EN-IT e INT-EN-ES per velocità (parole al minuto).

SPEED	ORG-EN	INT-EN-IT	INT-EN-ES
high > 160 w/m	34	4	12
medium 130-160 w/m	36	49	33
low < 130	11	28	36

Gli interpreti italiani rappresentati in questo sottocorpus hanno emesso i loro TA a una velocità media di 123,7 parole al minuto, mentre gli interpreti spagnoli si attestano a 137 parole al minuto. Quest'ultimo dato è sempre inferiore alla velocità dei rispettivi TP inglesi, ma è al contempo superiore a quanto è stato registrato per i colleghi italiani alle prese con gli stessi TP. In entrambi i sottocorpora di TA il numero di parole è decisamente inferiore rispetto al numero di parole totale dei corrispondenti TP (-6940 parole per la cabina italiana; -4639 parole per la cabina spagnola).

3.3.2.3 Sottocorpora INT-ES-IT e INT-ES-EN

Nel sottocorpus di discorsi in spagnolo sono state considerate le cabine di italiano e di inglese. La distribuzione di genere sulla base delle voci percepite negli interventi registrati presenta un quadro totalmente diverso per le due cabine. In italiano si sono ascoltate solamente voci femminili, mentre per la cabina inglese si riscontra una prevalenza della voce maschile:

Tabella 3.33 Numero di interventi interpretati da interpreti uomini o donne dal sottocorpus ORG-ES.

GENDER	INT-ES-IT	INT-ES-EN
M	**0**	**16**
F	**21**	**5**

Per quanto riguarda la lunghezza delle rese in termini di parole non sembrano esserci cambiamenti apprezzabili tra le due cabine rispetto ai discorsi originali:

Tabella 3.34 Numero di interventi in INT-ES-IT e INT-ES-EN per lunghezza (numero di parole).

TEXT LENGTH	ORG-ES	INT-ES-IT	INT-ES-EN
long > 1000	4	2	3
Medium 301-1000	10	9	10
short < 300	7	10	8

Anche in questi sottocorpora il parametro che presenta una variazione maggiore riguarda la velocità misurata in parole al minuto. Nuovamente si riscontra una diminuzione complessiva della velocità, anche se per la cabina italiana questo è valido rispetto ai discorsi originali segnalati come molto veloci. La velocità media (tra 130 e 160 parole al minuto) sembra essere stata mantenuta, pertanto non varia di molto il dato sulla velocità bassa. Un quadro diverso è invece offerto dalla cabina inglese che registra una diminuzione complessiva della velocità di emissione:

Tabella 3.35 Numero di interventi in INT-ES-IT e INT-ES-EN per velocità (parole al minuto).

SPEED	ORG-ES	INT-ES-IT	INT-ES-EN
high > 160 w/m	10	5	4
medium 130-160 w/m	7	13	9
low < 130	4	3	8

Il calcolo della media per la velocità di emissione dei TA è di 124,5 parole al minuto per la cabina italiana e di 136,2 parole al minuto per la cabina inglese. Entrambi i sottocorpora di TA presentano un numero totale di parole inferiore a quanto si può rilevare nel sottocorpus di TP spagnoli (-1.573 per la cabina italiana e -1.411 parole per la cabina inglese). Questa sembrerebbe essere una tendenza generale in tutto il corpus, con la sola eccezione del gruppo di TA spagnoli che sono stati prodotti a partire da discorsi originali italiani.

Capitolo 4

L'Archivio Multimediale e il Corpus DIRSI

In questo capitolo sono descritti l'archivio multimediale e il corpus elettronico DIRSI-C. Il primo è costituito da tutti i materiali raccolti in quattordici convegni presenziati da chi ha svolto il presente studio in qualità di *practisearcher* (solo in un caso, il convegno ELSA, abbiamo assistito all'evento comunicativo come semplici ricercatori-osservatori e non anche come parte dell'*equipe* di interpreti ingaggiati). I materiali comprendono la registrazione audio (in un solo caso anche video) dei TP, la registrazione audio dei TA, i documenti di supporto alle relazioni presentate dai conferenzieri (presentazioni power point, stampati), i programmi di ciascun evento e, solo per una parte dei materiali, le trascrizioni. Il secondo, il Corpus DIRSI-C, è lo strumento di ricerca vero e proprio, risultante dall'applicazione concreta del *corpus-based approach* allo studio di alcune parti di tre convegni tra tutti quelli inclusi in archivio (nello specifico, si tratta dei convegni CFF4, ELSA e CFF5, cfr. Tabella 4.1).

Di nuovo, le diverse tappe nella realizzazione dei corpora di interpretazione illustrate nel primo capitolo saranno riprese una ad una, al fine di fornire una rendicontazione puntuale e dettagliata di come sono state affrontate sul campo. Come già è stato ribadito più volte, la natura dei dati racchiusi in quest'altro corpus ha comportato notevoli differenze rispetto all'esperienza di ricerca maturata con EPIC, che ne ha comunque ispirato fortemente la metodologia e l'impostazione generale. L'intero capitolo potrebbe essere visto come una sorta di "diario di bordo", in cui la "narrazione" di questa esperienza pratica concorre a completare il precedente quadro teorico, fornendo tante conferme, ma anche numerose soluzioni alternative a quelle discusse o ipotizzate.

Nonostante l'ordine lineare che sarà seguito per presentare i vari punti illustrati nelle prossime sezioni, è doveroso specificare che l'andamento di tutte le fasi di questa ricerca è stato prevalentemente ciclico, con una continua

ridefinizione e rivisitazione di quanto si è andato realizzando di volta in volta.[1] Fatta questa precisazione, non si peccherà di incoerenza se il primo punto affrontato riguarda la creazione dell'archivio multimediale DIRSI, prima ancora di discutere le principali questioni pertinenti alla tappa dedicata al *corpus design*. L'archivio e il corpus sono due entità strettamente correlate (l'uno è "figlio" dell'altro), ma separate e distinte, se non altro per via del livello di trattamento dei dati affinché essi siano adoperabili nel corpus (e non solo come parte integrante dell'archivio).

Prima di illustrare in dettaglio tutte le fasi della creazione dell'archivio e del corpus, presentiamo sinteticamente il quadro generale dei dati raccolti, così da poter comprendere al meglio i riferimenti alle azioni specifiche della loro raccolta, catalogazione e successiva elaborazione.

I diversi convegni immagazzinati nell'archivio multimediale DIRSI sono stati tenuti tutti in Italia, tra il 2006 e il 2010, nelle città (o località in provincia) di Verona (PTE, CFF4, CFF5, CFF7, CFCARE), Bologna (HIST), Vicenza (BIRD), Terni (ML10, TICCIH E TICCIH-AG, EDLESI, STEELT), Cesena (ELSA) e Venezia (DAYSG). Tutte le sigle utilizzate per indicare i vari convegni sono state concepite per agevolare e snellire il riferimento a ogni singolo evento nel corso della presente trattazione. I convegni le cui celle risultano evidenziate nella Tabella 4.1 (CFF4, ELSA e CFF5) sono i tre eventi selezionati per essere inclusi nel corpus elettronico DIRSI-C (rec. = registrazione):

[1] Questo vale anche per il progetto di ricerca EPIC ed è in linea con quanto indicato da Biber (1993, p. 256) secondo cui «the compilation of a representative corpus should proceed in a cyclical fashion». Infatti «the design of a representative corpus is not truly finalized until the corpus is completed, and analyses of the parameters of variation are required throughout the process of corpus development in order to fine-tune the Representativeness of the resulting collection of texts» (*ibid.*).

Tabella 4.1 Elenco dei convegni contenuti nell'archivio DIRSI.

n.	Data rec.	Titolo principale dell'evento	Rif.	Totale rec.*	Ambito
1	22.03.2006	*Accessibility and Safety for All*.	**PTE**	214'	Sicurezza – Assistenza sociosanitaria
2	28.04.2006	Il nemico in politica. La delegittimazione dell'avversario nell'Europa Contemporanea nei secoli XIX e XX.	**HIST**	249'	Storia
3	20.05.2006	IV Seminario di Primavera. Progressi recenti e sviluppi futuri nella ricerca sulla fibrosi cistica: diabete, nutrizione, comunicazione via internet.	**CFF4**	310'	Medicina
4	27.05.2006	*Meeting on Rare Diseases. Genetic Therapies*.	**BIRD**	430	Medicina
5	16.06.2006	Financial development and savings in the growth process. A Schumpeterian approach.	**ML10**	100'	Economia
6	16.09.2006	TICCIH 2006 (sessione A) Patrimonio industriale e trasformazioni urbane.	**TICCIH**	400'	Archeologia industriale
7	20.09.2006	TICCIH 2006 Assemblea generale.	**TICCIH-AG**	90'	Gestione associativa
8	19.10.2006	Partecipazione e partnership nelle politiche locali a sostegno degli anziani non autosufficienti e dei loro famigliari.	**ELSA**	150'	Assistenza sociosanitaria
9	25-26-27.10.2006	*Day surgery* e *day services*: come realizzare il progetto di *day surgery*.	**DAYSG**	838'	Assistenza sociosanitaria
10	02.12.2006	*Equality and Diversity Learning in the European Steel Industry* (EDLESI).	**EDLESI**	180'	Pari opportunità
11	11.05.2007	V Seminario di Primavera. Progressi recenti e sviluppi futuri nella ricerca sulla fibrosi cistica: cosa cambia in FC, farmacoterapia del difetto di base, progressi nel trapianto polmonare FC.	**CFF5**	307'	Medicina
12	06.02.2009	Steel-Town 2009	**STEELT**	344'	Siderurgia – Urbanistica
13	16.05.2009	VII Seminario di Primavera. Progressi recenti e sviluppi futuri nella ricerca sulla fibrosi cistica: il registro europeo dei malati FC; le reti nordamericana ed europea per lo sviluppo di terapie FC; riflessioni di un malato sulla ricerca FC	**CFF7**	370'	Medicina
14	06.03.2010	*Meet the Team – Adult care in cystic fibrosis*. Assistenza al paziente adulto con fibrosi cistica: l'esperienza di un centro adulti europeo.	**CFCARE**	320'	Assistenza sociosanitaria

*Il minutaggio totale delle registrazioni è approssimativo e si riferisce al *floor*. Solo in due casi (convegni 6 e 7) il calcolo è basato sulla durata delle registrazioni del TA.

Per quanto sia stato possibile ottenere un buon numero di registrazioni da quando è stato avviato il progetto di ricerca, l'impegno richiesto nella realizzazione del corpus elettronico è stato tale per cui la selezione è stata limitata ai tre convegni indicati. Ad ogni modo, come si può cogliere a vista d'occhio osservando la Tabella 4.1, in totale vi sono ben otto convegni inerenti a

temi medici o legati all'assistenza sociosanitaria, per cui le possibilità di espansione di DIRSI-C sono più che plausibili. In particolare, la serie dei convegni CFF (CFF4, CFF5, CFF7, con la probabile aggiunta di altre edizioni future) potrebbe portare anche alla creazione di un corpus di interpretazione diacronico dalle caratteristiche uniche: si tratterebbe dello stesso tipo di evento (un convegno medico-divulgativo rivolto a personale medico, pazienti e familiari, per cui anche la comunità linguistica di riferimento e la sua diacultura rimarrebbero costanti), sullo stesso tema generale (la fibrosi cistica), con gli stessi interpreti (solamente in CFF5 è presente un'interprete che non partecipa invece a CFF4 e CFF7). A questi si potrebbe anche aggiungere il convegno CFCARE (pensato però specificatamente per il personale medico-sanitario), in cui sono discussi sempre temi riguardanti la fibrosi cistica, con la presenza della stessa coppia di interpreti ingaggiati in occasione dei vari convegni della serie CFF.

Un altro dato interessante che si può estrapolare osservando la Tabella 4.1, è la differenza tra i diversi convegni in termini di durata di tempo: si passa da un minimo di novanta-cento minuti (meno di due ore) con ML10 e TICCIH-AG fino a un massimo di quasi 14 ore totali per DAYSG (distribuite su tre giorni consecutivi). Esclusi questi casi più "estremi", i rimanenti convegni presentano una durata piuttosto simile (la media è di circa cinque ore); nel dettaglio, vi è un primo gruppo la cui durata globale non supera, o supera di poco le quattro ore (ELSA, EDLESI, PTE e HIST), mentre un secondo gruppo spazia dalle cinque alle sette ore di lavoro (in ordine crescente CFF5, CFF4, CFCARE, STEELT, CFF7, TICCIH e BIRD). Si noti che questo dato è calcolato in base alla durata delle registrazioni ottenute, pertanto è una rappresentazione in difetto della realtà. Gli eventi reali, così come si sono sviluppati nella loro globalità, hanno tutti avuto una durata superiore a quanto abbiamo indicato. Il dato a nostra disposizione è comunque utile, in quanto misura l'impegno di lavoro effettivo degli interpreti ai fini dell'erogazione del servizio di interpretazione simultanea. Partendo da un'altra prospettiva, il tempo di cui stiamo rendendo conto è il tempo occupato dagli eventi linguistici ratificati che sono stati prodotti durante lo svolgimento della situazione comunicativa, con esclusione dei momenti di pausa previsti dal programma, durante i quali la registrazione è sempre stata sospesa. A grandi linee, le registrazioni comprendono gli eventi linguistici prodotti nel corso delle varie sessioni (sessioni di apertura, di lavoro e di chiusura), nonché dei dibattiti e delle tavole rotonde, per un totale di circa 70 ore di materiale registrato (a cui corrispondono altrettante ore di registrazione dei TA). Come sarà spiegato approfonditamente più avanti, tuttavia, il corpus vero e proprio contiene solamente alcuni tipi di eventi linguistici tra tutti quelli che abbiamo riscontrato nelle registrazioni. Questa ulteriore azione di selezione è stata necessaria

prevalentemente ai sensi della rappresentatività dell'oggetto di studio, ma anche per esigenze pratiche di completamento del lavoro.

4.1 Impostazione dell'Archivio Multimediale

Così come è avvenuto per EPIC, prima ancora di iniziare la raccolta dei dati con cui realizzare un nuovo corpus di interpretazione, si pone da subito il problema di come tali dati possano essere gestiti nel modo più efficiente. In generale, la tecnologia digitale e, in particolare, l'informatica consentono di "trasportare" e realizzare copie dei dati raccolti con estrema facilità, utilizzando non solo i computer ma anche *hard disk* esterni e portatili sempre più capienti. Nel caso del progetto DIRSI, per l'acquisizione del TP abbiamo avuto a disposizione un computer portatile con una capacità di memoria piuttosto limitata (10GB). Pertanto, è stata necessaria fin da subito un'integrazione, apportata con l'uso di un *hard disk* esterno portatile molto più capiente (240GB), dove poter trasferire tutte le registrazioni raccolte.[2]

Per risalire facilmente a quali convegni corrispondono i singoli file delle registrazioni, ad ogni file è stato assegnato un nome-codice rispettando una precisa sequenza di informazioni. Tale sistema è stato utilizzato anche per i materiali EPIC (§3.1) ed è stato qui adattato alle caratteristiche dei materiali raccolti per DIRSI.

La sequenza di informazioni contenute nei nomi di ogni singolo file di registrazione è la seguente:

1.	nome del corpus	DIRSI
2.	data	YYYY-MM-DD
3.	sigla della città	(VR, FC, ecc.)
4.	sigla del convegno	(CFF4, CFF5, ELSA, ecc.)
5.	numero progressivo	00
6.	tipo di testo (originale o interpretazione)	ORG | INT

Ad esempio, per il primo convegno sulla fibrosi cistica (CFF4, tenuto a Verona) abbiamo immagazzinato sei file audio, corrispondenti alla registrazione dei TP e dei TA dell'intero convegno. In altri termini, sono stati ottenuti tre file separati

[2] I file audio sono stati salvati nel formato .WAV, per cui due ore circa di registrazione occupano quasi 700MG di spazio su disco. Si consideri che se all'inizio del progetto, nel 2006, le schede di memoria *flash memory* disponibili nel mercato arrivavano a un massimo di 70-100 MG, dopo solo quattro anni se ne trovano comunemente con una capienza attorno ai 2 GB e oltre.

per i TP e tre file separati per i TA. La registrazione dell'intero evento risulta suddivisa in tre parti sulla base della strutturazione del convegno in tre sessioni, separate da due momenti di pausa programmati. Durante tali pause, è stato possibile fermare la registrazione, salvare tutti i file e produrne una copia che è stata subito trasferita nel disco esterno. Di seguito sono riportati i nomi dei sei file così ottenuti:

DIRSI-2006-05-20-VR-CFF4-01-ORG.wav
DIRSI-2006-05-20-VR-CFF4-01-INT.wav

DIRSI-2006-05-20-VR-CFF4-02-ORG.wav
DIRSI-2006-05-20-VR-CFF4-02-INT.wav

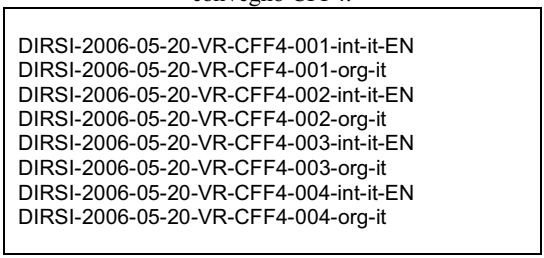
DIRSI-2006-05-20-VR-CFF4-03-ORG.wav
DIRSI-2006-05-20-VR-CFF4-03-INT.wav

Da questi file, corrispondenti a lunghi brani registrati durante il convegno, sono state estratte le clip individuali di ogni singolo evento linguistico di nostro interesse. Il nome di ogni clip, a sua volta, è stato composto secondo lo stesso sistema di denominazione, con l'aggiunta di altri due elementi informativi. Essi riguardano la lingua del singolo intervento (nel caso si tratti di un TP), indicata con il relativo codice linguistico ("it" per l'italiano e "en" per l'inglese), nonché la direzione linguistica e la direzionalità (nel caso si tratti di un TA). La direzione linguistica è espressa dall'ordine in cui compaiono i due codici linguistici; di questi, il codice in caratteri maiuscoli corrisponde alla lingua A, mentre il codice in caratteri minuscoli corrisponde alla lingua B (eventualmente alla lingua C) dell'interprete. Il risultante elenco degli elementi informativi usati per denominare le clip individuali è il seguente:

1.	nome del corpus	DIRSI
2.	data	YYYY-MM-DD
3.	sigla della città	(VR, FC, ecc.)
4.	sigla del convegno	(CFF4, CFF5, ELSA, ecc.)
5.	numero progressivo	000
6.	tipo di testo (originale o interpretazione)	org | int
7.	lingua (per i TP)	it | en
8.	direzione linguistica e direzionalità (per i TA)	int-en-IT | int-IT-en int EN-it | int-it-EN

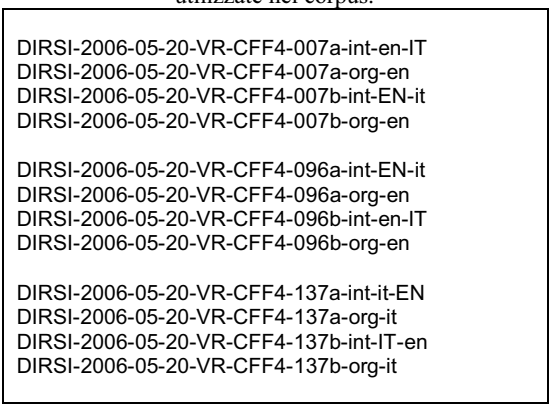
L'immagine sotto riportata (Figura 4.1) è un esempio di come appaiono nell'insieme alcuni dei file raccolti per lo stesso convegno (CFF4):

Figura 4.1 Esempio di denominazione delle clip ottenute dalle registrazioni integrali del convegno CFF4.



Tuttavia, nel caso di singoli eventi linguistici caratterizzati da una durata tale da costringere gli interpreti ad alternarsi, l'intero TP è stato diviso nelle due parti corrispondenti al TA prodotto da ciascun interprete, in quanto il corpus è strutturato sia a partire da un punto di osservazione esterno, quello dell'analista, sia da un punto di osservazione interno, ovvero quello degli interpeti. Al fine di poter cogliere anche questa particolarità da una semplice osservazione dei file salvati in archivio, è stata aggiunta una lettera al numero progressivo delle clip interessate da tale suddivisione. La seguente immagine (Figura 4.2) mostra la denominazione delle clip ottenute dal convegno CFF4 e utilizzate in DIRSI-C, interessate dalla suddivisione del turno di lavoro degli interpreti:

Figura 4.2 Visualizzazione della cartella contenente tutte le clip ottenute dal convegno CFF4 e utilizzate nel corpus.



Nell'immagine sopra riportata si può notare che i codici numerici assegnati ai singoli file non seguono un ordine sequenziale: da 009 si passa a 011, così come da 012 si passa a 093. Questo è dovuto alla selezione, effettuata a monte, del tipo di dati che si ritiene siano rappresentativi all'interno del corpus, in base a criteri frutto di riflessone teorica. A differenza di EPIC, infatti, le lingue di lavoro dei convegni registrati coincidono con le lingue di interesse all'interno del progetto di ricerca (italiano e inglese), pertanto tutta la registrazione dell'intero evento comunicativo sarebbe da includere nel corpus. Tuttavia, le stesse esigenze di archiviazione e catalogazione dei materiali raccolti hanno alimentato la necessità di classificare i dati attraverso parametri che rendessero conto di alcuni degli aspetti fondamentali insiti nell'evento comunicativo in questione, quali le sezioni o episodi di cui si compone la sua struttura, il ruolo comunicativo dei partecipanti e le caratteristiche degli eventi linguistici da loro prodotti (e mediati dagli interpreti).

Per motivi di spazio, non è possibile illustrare in dettaglio il percorso seguito al fine di giungere alla classificazione adottata. Ad ogni modo, vale la pena precisare che tale approccio è stato orientato da altre discipline che da sempre si occupano di studiare e analizzare diverse situazioni comunicative, quali la sociolinguistica (Berruto 1997, Goffman 1981), la linguistica antropologica e l'etnografia della comunicazione (Duranti 1997, Hymes 1980), l'analisi della conversazione e l'analisi del discorso (Hutchby & Wooffitt 1999; Heritage 1995, 1997; Gavioli 1999; Brown & Yule 1986; Shiffrin et al. 2001; Jaworski & Coupland 1999). Tutte queste fonti hanno fornito gli strumenti adatti ad integrare i contributi di coloro che già si erano occupati di trattare la conferenza mediata da interpreti come una situazione comunicativa (tra cui Namy 1978; Pöchhacker 1992, 1994a; Riccardi 1995, 2003; Russo 1999), così come hanno consentito di attingere dalle riflessioni di coloro che invece si sono concentrati sulla stessa situazione comunicativa, ma senza che fosse mediata da interpreti (tra gli atri, Shalom 1995, Räisänen 1999, Ventola et al. 2002). A conferma della validità di questo approccio multidisciplinare, con il quale l'osservazione è avvenuta a partire da due punti di vista (uno esterno e uno interno), uno dei "padri" della linguistica computazionale ha affermato che «The specification of a corpus […] is hardly a job for linguists at all, but more appropriate to the sociology of culture» (Sinclair 1991, p. 13).

Sulla base di tale presupposto e considerando il nostro oggetto di studio (vale a dire i convegni internazionali mediati da interpreti simultanei), abbiamo potuto constatare che tali eventi comunicativi presentano una macrostruttura in *fasi* delimitate temporalmente, corrispondenti a momenti che hanno luogo prima (fase precongressuale), durante (il convegno vero e proprio) e dopo l'evento comunicativo stesso (eventuale pubblicazione degli atti, ecc.).

Nella loro forma più semplice, tali eventi corrispondono alla conferenza intesa come evento, organizzato per consentire a un solo oratore o ospite invitato

di rivolgersi a un pubblico (di tenere, cioè, una conferenza nel senso italiano di discorso o presentazione). In realtà, anche in questi casi, difficilmente si avrà una sola persona che inizia e conclude un discorso di una certa lunghezza all'interno di un contesto vuoto. In genere, la prassi convegnistica vuole che l'oratore sia presentato da una persona che funge da moderatore (Shalom 1995). Questi assolve anche altre funzioni, soprattutto in riferimento alla gestione dei tempi e della facoltà di parola, nonché dell'eventuale dibattito che spesso segue l'intervento del conferenziere. Il moderatore, infine, ha anche il compito di concludere l'evento, annunciando a tutti i convenuti, in un certo senso, il permesso di lasciare il luogo in cui si erano riuniti, nonché di uscire dai ruoli comunicativi che si erano stabiliti in virtù dell'evento a cui stavano partecipando. È chiaro, quindi, come la stessa conferenza-convegno, come unità minima appena presentata, sia costituita da tante azioni comunicative, realizzate da diversi partecipanti (Ventola et al. 2002).

A partire da questa unità minima, nella quale si registra la presenza di un intervento principale (cioè una sola conferenza nel senso, come abbiamo spiegato, di discorso orale, relazione o presentazione), e di altri eventuali interventi, cioè eventi linguistici che vi ruotano attorno, è possibile arrivare a strutture più articolate. In esse, sono presenti più di un intervento principale, oltre a tutti gli altri interventi "accessori" che li accompagnano. A seconda del numero dei discorsi, della loro durata e degli oratori coinvolti, l'evento conferenza può articolarsi in maniera più o meno complessa, arrivando anche a strutturarsi in diversi momenti (sezioni o episodi) altamente ritualizzati e dotati di una certa autosufficienza rispetto all'architettura generale dell'evento, ovvero le *sessioni*. Qualora si abbia una strutturazione in più sessioni, per riferirsi a questi eventi comunicativi si utilizzano normalmente i termini "convegno" o "congresso", oltre a numerose altre diciture con sfumature di significato leggermente diverso (Russo 1999). Per il progetto DIRSI abbiamo stabilito che gli eventi comunicativi raccolti siano da considerare come "convegni internazionali".[3]

All'interno di un convegno, la *sessione di lavoro* o *tematica* seguita da eventuale dibattito è ripetibile più volte, in successione e anche parallelamente a

[3] Con il termine "convegno" si evita l'ambiguità che si avrebbe invece con il termine "conferenza", in quanto il significato di quest'ultimo rimanda sia all'evento comunicativo globale, sia a un «institutionalized extended holding of the floor in which one speaker imparts his views on a subject, these thoughts comprising what can be called his "text"» (Goffman 1981, p. 165). Riccardi (2009, p. 361) suggerisce una distinzione tra convegni bilaterali e convegni internazionali, a seconda del numero di lingue coinvolte. Nei convegni raccolti in DIRSI le lingue sono sempre e solo l'italiano e l'inglese, dunque si tratta di convegni bilaterali. Abbiamo preferito comunque la dicitura "convegni internazionali" poiché la presenza della lingua inglese non presuppone la partecipazione esclusiva di parlanti madrelingua, anzi il più delle volte sono presenti persone provenienti da diversi paesi non anglofoni (§4).

sessioni dello stesso tipo.[4] Ad esse, sono talvolta abbinate anche altre sessioni che segnano l'inizio e la fine dell'evento stesso nella sua globalità. Si tratta della *sessione di apertura* e della *sessione di chiusura* del convegno. Questi momenti sono di norma chiaramente identificabili nel programma come momenti "autonomi" e possono contenere al loro interno *eventi linguistici* di varia natura, quali le formule di benvenuto, oppure gli avvisi, le presentazioni plenarie non incluse nelle sessioni scientifiche o di lavoro, ecc. Si tratta, in pratica, dei momenti iniziali e conclusivi dell'intero evento (cioè della situazione comunicativa) nella sua complessità, da non confondere con gli eventi linguistici corrispondenti agli interventi di apertura e di chiusura delle singole sessioni di lavoro o delle altre sessioni tra quelle individuate prima. Tra le sessioni di lavoro, oltre alle *sessioni di presentazione*, ne esistono altre con formati interazionali diversi, quali la *tavola rotonda*, la sessione *poster* e la *discussione* o *dibattito*, anche se quest'ultimo è da ritenere subordinato alle sessioni di presentazione (Shalom 2002, p. 55). Infine, alle sessioni di lavoro si accompagnano le *sessioni sociali*, ovvero altri momenti appartenenti alla conferenza-convegno, ma non strettamente legati alla trattazione dei temi per i quali i delegati si sono appositamente incontrati (nelle modalità previste nel corso delle altre sessioni e con la relativa gestione della facoltà di parola). Lo sono, per esempio, i momenti di pausa tra una sessione e l'altra, i pranzi e le cene dette per l'appunto "sociali", le eventuali escursioni e così via. Tutte le tipologie di sessioni, con le caratteristiche precipue dei partecipanti e dei relativi eventi linguistici che sono stati individuati, sono riassunti nella Tabella 4.2. Essa contiene l'intero elenco delle caratteristiche fondamentali che abbiamo potuto rilevare a livello teorico e attraverso l'analisi dei programmi di tutti i convegni registrati, assieme ad altri dati che partecipano alla definizione della "identità" di ogni singolo evento linguistico ratificato, prodotto all'interno del convegno. È l'evento linguistico, infatti, l'unità minima a cui intendiamo applicare il nostro sistema di classificazione. La risultante serie di parametri di classificazione costituisce la base da cui, successivamente, sarà tratta una tassonomia funzionale all'inserimento dei dati nel corpus elettronico DIRSI-C (§4.2.3.3). Come vedremo, l'applicazione del *corpus-based approach* ai materiali a nostra disposizione comporterà un'ulteriore integrazione dei parametri, nonché la determinazione di valori soglia per alcuni di essi, così da poterli attribuire efficacemente ai singoli eventi linguistici rappresentati in DIRSI-C.

[4] Questo tipo di rappresentazione è in linea con lo sviluppo orizzontale e verticale descritto da Pöchhacker (1994b).

Tabella 4.2 Sintesi generale dei parametri di classificazione applicabili agli eventi linguistici ratificati nel convegno.

conference session:	opening presentation (paper) presentation (plenary) discussion round table or panel poster closing
speech event:	opening remarks paper lecture floor allocation procedure housekeeping announcements question answer comment closing remarks
duration:	short medium long
speech length:	short medium long
speed:	low medium high
speech delivery:	impromptu read mixed
audio visual support:	yes no
conference participant:	initiator organizer sponsor chair discussant or respondent presenter lecturer audience interpreter
language:	it en
native speaker:	yes no
directionality:	A B C
materials provided to interpreters:	in advance on the spot none

Tornando alla strutturazione dell'archivio multimediale, per poter avere un quadro generale dei dati raccolti e monitorare la produzione di clip e trascrizioni, è stato approntato un foglio di lavoro Excel apposito, strutturato in base ad alcuni dei parametri descritti precedentemente. Attraverso l'applicazione di filtri automatici, questo archivio informatizzato consente di calcolare velocemente il numero di dati a disposizione che presentano determinate caratteristiche. Le caratteristiche particolari che sono state incluse per classificare i dati in tale strumento operativo sono le seguenti:

- numero di parole TP
- numero di parole TA
- denominazione clip
- sigla convegno
- ruolo partecipante
- nome partecipante
- paese
- sessione
- categoria di evento linguistico
- durata in secondi
- lingua
- sigla interprete
- direzionalità.

Oltre alle informazioni ricavabili dalle varie voci incluse nell'archivio, l'impiego di diversi colori permette di differenziare e captare velocemente ulteriori dettagli, quali lo stato delle trascrizioni (bozza, revisione, definitiva), il tipo di evento linguistico (incluso o escluso dal corpus) e il convegno stesso. Un esempio tratto da uno dei fogli di lavoro impostati a tale scopo è riportato nella Tabella 4.3:

Tabella 4.3 Estratto dell'archivio informatizzato DIRSI-MA.

parole TP	parole TA	file audio	sigla convegno	ruolo partecipante	nome	paese prov.	sessione	evento linguistico	durata in secondi	lingua TP	interprete	direzionalità
312	359	DIRSI-2006-05-20-VR-CFF4-001-org-it	CFF4	ORGANIZER	Mastella	IT	**OPENING**	opening remarks	204	IT	UK-01	A
110	261	DIRSI-2006-05-20-VR-CFF4-002-org-it	CFF4	SPONSOR	Braggion	IT	**OPENING**	opening remarks	113	IT	UK-01	A
6	9	DIRSI-2006-05-20-VR-CFF4-003-org-it	CFF4	ORGANIZER	Mastella	IT	**OPENING**	floor allocation	6	IT	UK-01	A
364	435	DIRSI-2006-05-20-VR-CFF4-004-org-it	CFF4	SPONSOR	Ricciardi	IT	**OPENING**	opening remarks	222	IT	UK-01	A
53	57	DIRSI-2006-05-20-VR-CFF4-005-org-it	CFF4	ORGANIZER	Mastella	IT	**OPENING**	floor allocation	28	IT	UK-01	A
241	205	DIRSI-2006-05-20-VR-CFF4-006-org-it	CFF4	CHAIR	Minicucci	IT	**PRESENTATION**	opening remarks	116	IT	UK-01	A
		DIRSI-2006-05-20-VR-CFF4-007-org-en	CFF4	P/L								
3258	3077	DIRSI-2006-05-20-VR-CFF4-007a-org-en	CFF4		Moran Antoinette	USA	**PRESENTATION**	presentation	1574	EN	UK-01	A
3017	2763	DIRSI-2006-05-20-VR-CFF4-007b-org-en	CFF4		Moran Antoinette	USA	**PRESENTATION**	presentation	1402	EN	IT-01	B

A partire dai dati a disposizione e sfruttando le relative informazioni usate per classificarli, sarebbe possibile allestire una vera e propria banca dati, con una interfaccia dotata di filtri di ricerca, nonché collegamenti ai file multimediali e alle trascrizioni. Strumenti simili già realizzati sono le banche dati e gli *speech repositories* descritti in precedenza (§1.2.1). La realizzazione di una banca dati del genere, contenente tutti i materiali raccolti nell'archivio multimediale DIRSI editati e trascritti, rientra negli sviluppi futuri del presente studio.

Ulteriori dettagli riguardanti tutti i materiali contenuti nell'archivio multimediale DIRSI saranno forniti anche dalla prossima sezione, a partire dalla quale l'attenzione sarà rivolta alla creazione di DIRSI-C. Va ribadito che il corpus è stato realizzato a partire da una selezione di eventi linguistici, tratti da tre convegni (CFF4, ELSA e CFF5) fra tutti quelli in archivio; le prime due tappe della creazione del corpus (*corpus design* e *data collection*, §1.3.1 e §1.3.2 rispettivamente) sono comuni a tutti i materiali, mentre le tappe successive riguardano i soli materiali selezionati a fare parte del corpus elettronico vero e proprio.

4.2 Creazione del Corpus

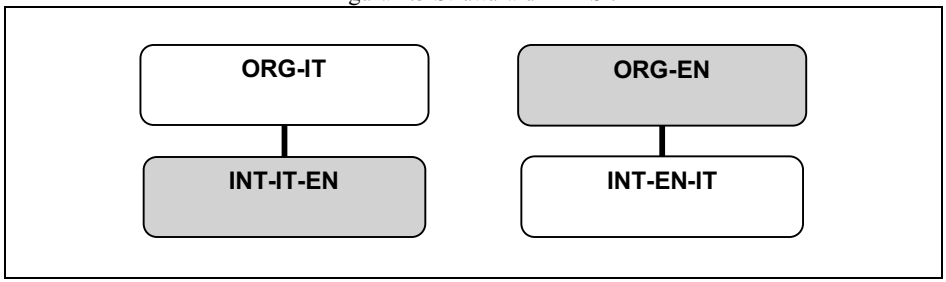
DIRSI-C è il corpus elettronico che è stato creato a partire da una selezione di dati appartenenti a tre dei quattordici convegni inclusi nel suo archivio multimediale. Il corpus rientra a pieno titolo nella categoria dei D.I.Y. (*do it yourself*) corpora proposta da McEnery et al. (2006, p. 71), in quanto è stato creato totalmente *ex novo*. Tuttavia, si deve riconoscere che una buona parte della metodologia utilizzata era già stata messa a fuoco durante la realizzazione del corpus EPIC.

4.2.1 Struttura e rappresentatività del corpus

DIRSI-C è un corpus bilingue (italiano e inglese) e parallelo, frutto cioè della raccolta di TP e TA nelle due lingue coinvolte, in entrambe le direzioni (e direzionalità). In altre parole, comprende TP in italiano con i relativi TA in inglese, e viceversa. I TA sono prodotti da interpreti professionisti in modalità simultanea (con cabina), i quali hanno lavorato sia verso la lingua A, sia verso la

lingua B.[1] La struttura globale del corpus è formata da quattro sottocorpora,[2] due con i TP e due con i TA, come rappresentato nella Figura 4.3:

Figura 4.3 Struttura di DIRSI.



(org = testo originale, TP; int = interpretazione, TA; IT = italiano; EN = inglese)

Come si vede chiaramente dalla rappresentazione schematica riportata nella Figura 4.3, DIRSI-C può essere utilizzato non solo come corpus parallelo, confrontando cioè i TP con i rispettivi TA, ma anche come corpus comparabile; ad esempio, si possono mettere a confronto i TP italiani con i TA, sempre italiani, che sono stati prodotti a partire dai TP inglesi. Inoltre, all'interno di ciascun sottocorpus, i dati possono essere ulteriormente filtrati sulla base degli attributi inclusi all'interno dell'*header* assegnato a ogni singola trascrizione (§4.2.3.3). Tali attributi non sono altro che le caratteristiche più rilevanti dei TP e dei TA, in base alle quali cui è stato possibile classificare i dati raccolti, con l'aggiunta di opportune integrazioni.

La questione della rappresentatività del corpus è stata affrontata con particolare intensità in due fasi dello svolgimento del progetto di ricerca, ovvero prima della raccolta dei dati e durante la trascrizione degli stessi. Prima di iniziare a raccogliere i dati effettuando le registrazioni, non avevamo ancora stabilito tutti i criteri di inclusione e di esclusione che sono stati effettivamente considerati. L'universo di campionamento da cui partire era senz'altro costituito dagli eventi comunicativi mediati da interpeti professionisti. Sapevamo di voler concentrare i nostri sforzi su un determinato tipo di interpretazione, ossia l'interpretazione di conferenza (il nostro oggetto di studio), con particolare riguardo alla modalità simultanea con cabina, svolta nell'ambito di convegni internazionali in Italia. Pertanto, la popolazione target di riferimento comprende tutti i soggetti che si possono considerare parte attiva del convegno vero e

[1] Dei sei interpreti coinvolti nella registrazione di tutto il materiale, solo uno è madrelingua inglese (tutti gli altri sono madrelingua italiani e hanno inglese come lingua B). Per una descrizione del profilo dei diversi interpreti, si veda più avanti (§4.3.1.1).

[2] Partendo da un'altra prospettiva, l'intero corpus potrebbe essere visto come la somma di tanti piccoli sottocorpora individuali per ogni convegno (DIRSI-C_CFF4 + DIRSI-C_ELSA + DIRSI-C_CFF5).

proprio, in particolare gli interpreti e gli organizzatori per accedere ai dati. Data la notoria difficoltà a ottenere la loro collaborazione (§1.3.2), l'approccio iniziale è stato aperto a qualsiasi proposta che rientrasse nella modalità e nel contesto prescelti, senza limitazioni di argomento o legate al profilo della comunità linguistica coinvolta. L'unico criterio di inclusione/esclusione ritenuto valido fin da subito riguarda la natura pubblica degli eventi (non soggetti a particolari restrizioni dovute alla riservatezza dei contenuti trattati) e la natura professionale del servizio di interpretazione, nel senso che saremmo andati alla ricerca solamente dell'operato di interpreti professionisti, ben inseriti nel mercato Italiano. Riguardo alle lingue, l'obiettivo (ispirazione) iniziale prevedeva la raccolta di dati da convegni con tre lingue di lavoro (italiano, inglese e spagnolo), in modo da riflettere fedelmente la struttura di EPIC. Tuttavia, si deve riconoscere che una tale configurazione linguistica limita notevolmente le dimensioni del nostro oggetto di studio, per non parlare delle sfide di tipo tecnologico e organizzativo insite nella gestione di una tale tipologia di raccolta. In definitiva, pur consapevoli della prevalenza della lingua inglese nel mercato dell'interpretazione di conferenze in Italia, siamo rimasti aperti a ogni possibilità che ci si fosse presentata, contemplando quindi tutte le combinazioni tra le lingue di lavoro di nostra competenza (italiano, inglese e spagnolo).

L'altra fase in cui la questione della rappresentatività si è riproposta con particolare vigore è la tappa della trascrizione dei dati. Essa ha coinciso con l'analisi dell'evento comunicativo dato dal convegno nella sua globalità, nel momento in cui siamo passati da un punto di osservazione esterno a un punto di osservazione interno. La trascrizione ha permesso di fissare la natura effimera del parlato ratificato durante il convegno; si è così materializzato un riferimento stabile, con il quale la strutturazione della situazione comunicativa si è palesata concretamente, mettendo in luce le sessioni e gli eventi linguistici di cui si compone. Questi elementi strutturali sono diventati la parte "viva" dell'oggetto di studio e della popolazione target precedentemente inquadrati, per cui è stato a partire da questa seconda matrice di dati che è stata effettuata un'ulteriore selezione al fine di realizzare il corpus. A ben vedere, tutte le sessioni e tutti gli eventi linguistici potrebbero essere inclusi di diritto nel corpus elettronico, in quanto manifestazioni comunicative della popolazione target, o meglio del "corpus di popolazione" ottenuto nel corso della nostra raccolta. Tuttavia, l'impegno richiesto dal compito della trascrizione in generale, ma ancor più le marcate differenze nel tipo di interazione in certe sessioni in particolare, ci hanno orientato verso la selezione solamente di alcuni tipi di sessione e di eventi linguistici da includere effettivamente nel corpus. Priorità è stata data alle sessioni di apertura, di presentazione e di chiusura dei convegni, i cui eventi linguistici sono stati interamente considerati. Diversamente, nelle sessioni di discussione e nelle eventuali tavole rotonde, sono stati selezionati solo alcuni

tipi di eventi linguistici, quali gli interventi di apertura, di chiusura e le eventuali conferenze-intervento o relazioni programmate in apertura o in chiusura di tutto l'evento. In questo modo, sono state escluse le parti di convegno che presentano un grado di interazione tra i partecipanti decisamente più alto rispetto a quanto avviene nelle rimanenti parti. Questo è dimostrato dal fatto che un maggior numero di casi di sovrapposizione è sempre stato riscontrato durante le sessioni di discussione, delineando di fatto due profili interazionali ben diversi tra loro e compatibili con la distinzione proposta da Hayashi (1996) tra *single conversational floor* e *multiple conversational floor*. Tra l'altro, il fenomeno della sovrapposizione dei parlanti non poteva essere gestito al meglio con il sistema di trascrizione adottato, rafforzando ulteriormente la nostra scelta ad escludere (per il momento) tale parte di dati. Ciò non toglie la possibilità di analizzare le sessioni di dibattito in futuro, integrandole in DIRSI-C, oppure costituendo un corpus apposito, seguendo un'impostazione che consenta di gestire al meglio la natura più dialogica che monologica degli eventi linguistici interessati.

4.2.2 Raccolta dei dati

Sulla base della definizione dell'universo di campionamento e della popolazione *target* impostata all'inizio del progetto, è stata messa a fuoco una "lista" di contatti potenzialmente utili alla raccolta dei materiali sul campo, ovvero è stato ipotizzato un *sampling frame* (§1.3.1.2). Come illustrato precedentemente, sono due le reti di contatti da considerare in un progetto CIS: i referenti responsabili dei TP e i referenti responsabili dei TA. Tuttavia, i primi non sempre si occupano dell'organizzazione del convegno; è ammissibile che gli organizzatori partecipino con eventi linguistici ratificati, per esempio in veste di relatori o moderatori, ma non risulta altrettanto vero il percorso inverso, in quanto sarebbe difficile partire dalla lista completa dei relatori e dei moderatori per individuare chi si occupa dell'organizzazione del convegno. Nella realtà dei fatti, tra i primi soggetti da contattare per avere accesso ai dati vi sono proprio gli organizzatori, con precedenza sui relatori (nonostante siano questi ultimi ad essere registrati). Analogamente, i referenti responsabili dei TA, ossia gli interpreti, di rado hanno voce in capitolo nell'organizzazione generale del convegno. Tuttavia, data la loro centralità, sia nella realizzazione dell'evento comunicativo mediato, sia nello svolgimento della ricerca, essi sono probabilmente l'anello migliore a cui agganciarsi per entrare in contatto con tutti gli altri soggetti coinvolti. Dopo tutto, quale che sia il tipo di ingaggio (da cliente diretto o da soggetti terzi, come i PCO), gli interpreti si interfacciano naturalmente con ognuna delle popolazioni incluse in tutte le fasi di cui si compone il macrosistema del convegno.

Avendo presenziato tutti gli eventi in qualità di *practisearcher* (con la sola eccezione del convegno ELSA), non è stato difficile individuare i soggetti a cui inoltrare le richieste di collaborazione. Ciononostante, va ammesso che questa duplice responsabilità (come interprete in servizio e come ricercatore sul campo) ha comportato uno sforzo considerevole nella gestione delle relazioni pubbliche e nel garantire il felice esito di tutte le attività previste.

4.2.2.1 Accessibilità

Dalla discussione condotta nel primo capitolo (§1.3.2.1), è emerso che al fine di accedere ai dati è possibile sfruttare canali sia esterni, sia interni all'evento comunicativo. Nel caso di eventi trasmessi in *streaming* o alla televisione e alla radio, così come con le banche dati multimediali in Rete, ci si avvicina ai dati da un canale esterno (come lo è stato per EPIC); dall'altra parte, nella ricerca sul campo, il più delle volte il canale di accesso è interno all'evento comunicativo stesso, e parte già dal contatto con i soggetti delle popolazioni *target* coinvolte sopra citate.

Nel caso specifico di DIRSI, nella fase iniziale dello studio è stata improntata una scheda informativa sul progetto di ricerca, al fine di divulgarne i contenuti a tutti i colleghi interpreti conosciuti personalmente e ai docenti di interpretazione della SSLMIT (con italiano, inglese e spagnolo come lingue di lavoro). Grazie al forte appoggio istituzionale dato dall'Università di afferenza dell'autore del presente studio, le prime reazioni furono più che incoraggianti. Tutti gli interpreti contatti direttamente (molti dei quali docenti in corsi universitari di interpretazione) si erano detti disponibili a fornire loro registrazioni e a contattare direttamente il *practisearcher* alla prima occasione. La realtà dei fatti ci mostra che alle parole non sempre sono seguiti i fatti, e la nostra politica fu quella di non insistere eccessivamente, in modo da non "bruciare il terreno" a possibili studi futuri. Fortunatamente, alcuni interpreti hanno dato seguito alla loro dichiarazione di disponibilità, acconsentendo a partecipare concretamente allo studio. Alcuni di loro si sono addirittura attivati in prima persona per fare da tramite e informare gli organizzatori dei convegni – una sorta di *inside champion*, come lo ha definito Pöchhacker (§1.3.2.2), che ha appianato ogni eventuale remora degli organizzatori e, conseguentemente, degli oratori partecipanti. Solo due interpreti furono contattati indirettamente, ossia attraverso altri colleghi o dallo stesso PCO. In un caso, l'interprete decise solo alla fine del convegno di non dare il proprio consenso (pur avendo fornito un servizio impeccabile e dicendosi d'accordo con i contenuti espressi nella scheda informativa sul progetto DIRSI che aveva ricevuto previamente). Nell'altro caso, l'interprete fu disponibile a fornire il proprio consenso, motivando però esplicitamente la propria decisione in virtù della sua fiducia nei confronti

dell'altra collega che aveva fatto da tramite e delle persone di sua conoscenza che stavano supervisionando la presente ricerca.

In totale, hanno collaborato al presente studio cinque interpreti di madrelingua italiana (compreso l'autore del presente studio) e un interprete di madrelingua inglese, con la garanzia dell'anonimato. Per questo motivo, i singoli soggetti interpreti sono indicati con i seguenti codici: IT-01, IT-02, IT-03, IT-04, IT-05 (i madrelingua italiani) e UK-01 (il madrelingua inglese). Maggiori dettagli su questi soggetti che hanno dato la loro preziosa disponibilità sono forniti più avanti (§4.3.1.1).

La partecipazione diretta dell'autore del presente studio come interprete in servizio in tutti i convegni raccolti (tranne uno) si espone sicuramente ad alcune critiche, specialmente per quel che riguarda l'analisi delle proprie prestazioni e la possibilità effettiva di raccogliere dati con un approccio osservazionale (si vedano le considerazioni precedentemente espresse a tal riguardo §1.3.2.3). D'altro canto, proprio grazie al coinvolgimento diretto dell'autore del presente studio come interprete-ricercatore è stato possibile avere un accesso relativamente facile ai dati, nonché conoscere appieno i dettagli organizzativi di ogni evento comunicativo e tutte quelle informazioni che, diversamente, dovrebbero essere estrapolate attraverso interviste, protocolli *think-aloud* e questionari. Tutti questi sono esempi di strumenti di non facile impiego e che non sempre consentono di ottenere le informazioni desiderate.

Dopo aver assicurato l'accesso ai dati da un canale interno all'evento comunicativo, sono due gli strumenti di cui è stato necessario munirsi per procedere alla raccolta effettiva dei dati: il consenso informato (di cui si è già discusso a livello teorico, §1.3.2.2), nonché le strumentazioni di ausilio alla registrazione e alla gestione delle registrazioni stesse (per le considerazioni teorico-pratiche generali si veda §1.3.2.3).

4.2.2.2 Consenso informato

Nel presente studio sono stati approntati due diversi documenti (riportati al termine di questa sezione, Figura 4.4 e Figura 4.5). Il primo è una scheda informativa, redatta allo scopo di illustrare ai potenziali partecipanti il progetto di ricerca e motivarli a partecipare attivamente. In questo documento, redatto su carta intestata, sono chiaramente espressi gli obiettivi generali e si fa leva sull'importanza di collaborare al progetto. Il secondo documento è il modulo vero e proprio utilizzato per il consenso informato. Qui il testo descrittivo è stato ridotto notevolmente, mettendo maggiormente in evidenza le condizioni d'uso dei dati raccolti:

> **Il materiale registrato sarà utilizzato solo ed esclusivamente per scopi accademici, escludendo fin d'ora qualsiasi tipo di impiego commerciale, nonché l'uso di tutti i materiali registrati per predisporre gli atti che saranno eventualmente a cura degli organizzatori.**

L'ultima parte della formula sopra riportata (evidenziata in grigio) è stata aggiunta dopo aver registrato il primo convegno raccolto (PTE, non utilizzato in DIRSI-C, ma disponibile nell'archivio DIRSI-MA), in quanto una simile precisazione fu espressamente richiesta dagli organizzatori. Curiosamente, in questo caso gli organizzatori lamentarono addirittura l'assenza di una videocamera per registrare, consapevoli del ruolo fondamentale della comunicazione non verbale.[3]

Si noti che nella formula considerata non sono presenti riferimenti all'anonimato delle persone. Dato che ci siamo concentrati su eventi di natura pubblica, si è ritenuto importante non dover effettuare cancellazioni successive nei dati originali del TP. Probabilmente, la stessa scelta non sarebbe riproponibile in contesti più sensibili o nel caso di riunioni "a porte chiuse", in cui le informazioni scambiate sarebbero di tipo riservato.

Un'ulteriore aggiunta effettuata a seguito dell'esperienza di registrazione del primo convegno riguarda la parte evidenziata nella seguente formula:

> **Acconsento alla registrazione degli interventi impegnandomi a informare i partecipanti:**

L'aggiunta alla formula sopra riportata è stata ritenuta di vitale importanza al fine di snellire il processo di reperimento del consenso. Nel primo convegno registrato, infatti, la richiesta di poter effettuare la registrazione fu trasmessa nella fase pre-congressuale (con l'invio della scheda informativa) inizialmente al PCO, subito dopo aver formalizzato l'ingaggio degli interpreti, e poi agli organizzatori e all'altra interprete. Come abbiamo già sottolineato, gli organizzatori si mostrarono estremamente disponibili, al punto da informare per posta elettronica tutti i conferenzieri in programma (da cui l'idea che potesse sempre essere l'organizzatore a garantire di farsi da tramite per informare tutti i partecipanti). Il giorno del convegno, il *practisearcher* prese immediatamente

[3] Gli organizzatori provenivano da una istituzione accademica che si occupa di ricerca ed erano consapevoli di tanti aspetti metodologici aventi un ruolo preponderante nella ricerca sul campo. Si tratta di un aspetto particolarmente interessante, in quanto è un'indicazione chiara del maggior grado di disponibilità presente in talune popolazioni linguistiche/diaculture rispetto ad altre. Queste potrebbero essere contattate direttamente, per avere accesso ai dati di nostro interesse, nel caso in cui organizzassero convegni o altri eventi mediati da interpreti. In tale eventualità, il *practisearcher* potrebbe mettere a disposizione la propria esperienza per fornire assistenza organizzativa, facendo in modo che tutti gli oratori e gli interpreti ingaggiati acconsentano alla registrazione già nella fase pre-congressuale.

contatto con la responsabile dell'organizzazione scientifica dell'evento e con tutte le persone che avrebbero ricoperto il ruolo di moderatore nelle varie sessioni previste. In questo modo, tutti gli oratori furono invitati a firmare il modulo del consenso informato poco prima di prendere la parola. Inoltre, un annuncio venne fatto a tutti i presenti un attimo prima di aprire i lavori. Nonostante questa situazione possa apparire ideale per l'alto livello di collaborazione e disponibilità, emerse comunque una serie di "intoppi" che ci hanno spinto ad aggiungere la formulazione sopra citata, in modo da dover ottenere concretamente solo una firma (da un responsabile dell'organizzazione). Tra gli "intoppi" che si verificarono, possiamo citare i seguenti tra i più esemplificativi:

- Alcuni oratori arrivarono tardi o poco prima di tenere il loro intervento. Non avevano avuto modo di sentire l'annuncio, ma probabilmente nemmeno di leggere il messaggio inviato dagli organizzatori che li informava della registrazione, oppure non se ne ricordavano in quel preciso istante. Conseguentemente, una volta preso posto al tavolo, non capivano perché dovessero firmare un foglio, pur avendo lì a fianco il moderatore che prontamente forniva le debite spiegazioni e pur potendo leggere velocemente il testo del modulo per il consenso davanti ai loro occhi. Fu necessario rincorrere letteralmente alcuni oratori durante le pause o al termine del convegno per far apporre la loro firma (ben felici di farlo!).
- Molte persone dal pubblico presero la parola durante i dibattiti, rendendo impossibile a chiunque di recuperare le firme di tutti senza disturbare lo svolgimento dei lavori.
- Alcuni partecipanti lasciarono la sala prima della conclusione dell'evento, rendendo al *practisearcher* impossibile l'ottenimento della loro firma (perché impegnato in cabina). Riteniamo che anche nel caso in cui si avesse un ricercatore completamente dedicato alla raccolta dei dati, il recupero del consenso dalle persone che abbandonassero l'evento prima del termine comporterebbe un certo disturbo alla prosecuzione dei lavori, così come una perdita di monitoraggio da parte del ricercatore.

A fronte di questi e altri ostacoli alla raccolta delle firme per il consenso informato, con l'aggiunta della formula sopra riportata è stato possibile gestire questo momento delicato con più serenità, senza annunci "intimidatori" all'inizio dei lavori e senza dover effettuare acrobazie per raggiungere ogni singola persona che avesse preso la parola nel corso del convegno.

Bisogna ricordare che oltre ad approntare tutto quanto è necessario per ottenere il consenso dai partecipanti, il *practisearcher* coinvolto a pieno titolo

nella raccolta deve avere anche il tempo utile per installare le strumentazioni di registrazione con il tecnico e di prendere posto in cabina con l'altro interprete per svolgere seriamente il proprio servizio. In tutto ciò, i momenti antecedenti l'inizio di un convegno possono rivelarsi complessivamente frenetici e il nervosismo è talvolta palpabile. Per questo, è fondamentale conoscere esattamente la sequenza di azioni da compiere e necessarie al buon esito della raccolta dei dati. Ciò è soprattutto vero per il *practisearcher*, ma lo è anche nel caso in cui il ricercatore non sia direttamente coinvolto come interprete in servizio durante il convegno oggetto di studio.[4]

Ovviamente, l'inclusione di questa formula nel modello di consenso informato implica che gli organizzatori siano informati con un buon anticipo, e che ricevano copia della scheda informativa e del modulo per il consenso informato il prima possibile. Ad ogni modo, anche se gli organizzatori non informassero effettivamente tutti i partecipanti (anche per i motivi rilevati precedentemente), questo metodo rimarrebbe valido fintanto che gli eventi in questione sono di natura pubblica, i cui contenuti non sono sottoposti a restrizioni dovute alla tutela dei diritti di autore o a questioni di riservatezza (ammesso che se ne faccia un uso legittimo, §1.3.2.2). Nella nostra esperienza, tutti gli organizzatori hanno mostrato grande disponibilità, specialmente laddove esisteva già un rapporto di stima professionale con almeno uno degli interpreti. In questi casi, i responsabili dell'organizzazione del convegno davano addirittura l'impressione che avrebbero firmato qualunque cosa, tanto era entusiastica la loro reazione nel poter collaborare alla ricerca. Solo in un caso si creò momentaneamente una situazione a dir poco imbarazzante: dopo aver fornito il consenso alla registrazione qualche giorno prima che si svolgesse il convegno, gli organizzatori ricontattarono gli interpreti per chiedere uno sconto sulla tariffa accordata, quale premio per il loro gesto di cooperazione e a cambio dei vantaggi professionali di cui l'autore del presente studio avrebbe beneficiato anche grazie alla registrazione in questione. Data l'indebita richiesta, gli organizzatori furono dapprima invitati dagli interpreti a valutare quanto fosse opportuno porre la questione in quei termini, dopodiché fu loro spiegato che la registrazione non era assolutamente indispensabile visto che molti altri si erano comportati diversamente. L'errore fu subito riparato e venne fornito il consenso a registrare senza balzelli o detrazioni ingiustificate.[5]

---

[4] A tal proposito, è consigliabile preparare una *check-list* con tutte le operazioni da effettuare nella sequenza esatta in cui devono essere realizzate. Con l'esperienza, tale strumento diventa meno indispensabile, ma per chi si appresta a raccogliere dati per la prima volta diventa un elemento di vitale importanza.

[5] Per ironia della sorte, durante la registrazione della prima parte di questo convegno si verificò un problema tecnico e i dati registrati per la prima sessione andarono irrimediabilmente persi. Gli stessi organizzatori offrirono una copia delle audiocassette con cui era stato registrato il convegno per uso interno (autonomamente rispetto alla registrazione effettuata dal *practisearcher*).

Tornando al testo del modulo per il consenso informato, la garanzia dell'anonimato è invece presente nel modello utilizzato per ottenere il consenso dagli interpreti:

> **Il materiale registrato sarà utilizzato solo ed esclusivamente per scopi accademici, nel rispetto dell'anonimato degli interpreti, escludendo fin d'ora qualsiasi tipo di impiego commerciale, nonché l'uso di tutti i materiali registrati per predisporre gli atti che saranno eventualmente a cura degli organizzatori.**

Dall'esperienza illustrata in questa sezione, si può constatare che la divulgazione degli obiettivi di una ricerca, assieme alla precisazione che non si intende studiare solo ed esclusivamente le anomalie del TA e alla garanzia dell'anonimato, non sono sempre sufficienti a convincere gli interpreti perché diano il proprio consenso alla registrazione e allo studio dei dati. A ben vedere, l'elemento essenziale e costante in questa delicata fase di accesso ai dati è costituito dalle relazioni professionali e personali preesistenti tra il ricercatore e i soggetti interpreti coinvolti. Quanto maggiore è l'esperienza e la maturazione professionale di un *practisearcher*, tanto maggiori saranno le possibilità di poter contare sulla collaborazione di colleghi altrettanto esperti. Nel nostro caso, il *practisearcher* era ancora all'inizio della sua carriera, per cui la fonte principale dei contatti a sua disposizione corrisponde al contesto accademico in cui si è formato come interprete e in cui è stata svolta la presente ricerca. In altri casi, colleghi che contano su un'esperienza più matura hanno sicuramente accesso a un ventaglio di potenziali collaboratori più ampio. A questo proposito, si vedano gli esempi a cui si è fatto riferimento precedentemente (§1.3.2) tra cui, in particolare, Monacelli (2009, p. 33) conferma che: «Access to participants was negotiated with interpreters with whom I have an in-group relationship».

Figura 4.4 Modello di consenso informato per la registrazione nel progetto DIRSI.

PROGETTO DI RICERCA *DIR-SI Corpus*
"Directionality in Simultaneous Interpreting CORPUS"
(corpus sulla direzionalità in interpretazione simultanea)

CONSENSO

Il *DIR-SI Corpus* è un progetto di ricerca sull'interpretazione di conferenza condotto dal dott. Claudio Bendazzoli come tesi di dottorato presso il Dipartimento di Studi Interdisciplinari su Traduzione, Lingue e Culture (SITLeC) dell'Università di Bologna con sede a Forlì.

L'obiettivo del progetto è studiare il linguaggio di oratori e dei loro interpreti simultanei che traducono anche verso la lingua straniera (lingua "B") e non solo verso la propria lingua materna (lingua "A").

Si intende pertanto realizzare un archivio multimediale contenente registrazioni (video o audio) e trascrizioni in formato elettronico del materiale oggetto di studio, per analizzarlo attraverso tecniche appartenenti alla linguistica dei corpora.

Il materiale registrato sarà utilizzato solo ed esclusivamente per scopi accademici (ricerca e didattica), escludendo fin d'ora qualsiasi tipo di impiego commerciale, nonché l'uso di tutti i materiali registrati per predisporre gli atti che saranno eventualmente a cura degli organizzatori.

Acconsento alla registrazione degli interventi impegnandomi a informare i partecipanti:

_____
_____

Per ulteriori informazioni potete contattare:

Claudio Bendazzoli

EMAIL cbendazzoli@sslmit.unibo.it
TEL. +39 0543 374727
FAX +39 0543 374717
CELL. +39 349 2240102

Figura 4.5 Scheda informativa usata congiuntamente al consenso informato per la registrazione nel progetto DIRSI.

PROGETTO DI RICERCA *DIR-SI Corpus*
"Directionality in Simultaneous Interpreting CORPUS"
(corpus sulla direzionalità in interpretazione simultanea)

Il *DIR-SI Corpus* è un progetto di ricerca sull'interpretazione di conferenza condotto dal dott. Claudio Bendazzoli come tesi di dottorato presso il Dipartimento di Studi Interdisciplinari su Traduzione, Lingue e Culture (SITLeC) dell'Università di Bologna con sede a Forlì.
L'obiettivo del progetto è studiare il linguaggio di oratori e dei loro interpreti simultanei che traducono anche verso la lingua straniera (lingua "B") e non solo verso la propria lingua materna (lingua "A"). Si intende pertanto realizzare un archivio multimediale contenente registrazioni (video o audio) e trascrizioni in formato elettronico del materiale oggetto di studio, per analizzarlo attraverso tecniche appartenenti alla linguistica dei corpora.

Per poter realizzare tale studio, è necessaria la collaborazione di più soggetti, quali gli interpreti, gli oratori e gli organizzatori dell'evento. In passato, la pratica di registrare un intervento o una traduzione per analizzarli ha raramente riscosso consensi, per timore che il materiale fosse diffuso indiscriminatamente o che si ponesse attenzione esclusivamente sulle anomalie del parlato, soprattutto sulle difficoltà degli interpreti. Queste riserve devono oggi essere ridimensionate, soprattutto in virtù della straordinaria evoluzione che ha caratterizzato gli studi sulla lingua parlata e di un approccio proattivo agli studi sull'interpretazione e sulla traduzione, con il quale si vogliono mettere in evidenza le strategie di mediazione culturale e linguistica, nonché sui miglioramenti nella fruibilità di un dato testo.

Con la presente si vuole portare a conoscenza dell'attivazione del suddetto progetto di ricerca la vostra organizzazione, con la richiesta di autorizzare il Dott. Bendazzoli Claudio a effettuare le necessarie registrazioni nel caso in cui anche gli oratori e gli interpreti interessati esprimano consenso favorevole.

Il Dott. Bendazzoli si occuperà personalmente della registrazione dei dati con l'ausilio di miniregistratori digitali e computer portatile, nonché della raccolta del consenso firmato da parte dei partecipanti, garantendo fin d'ora che i materiali registrati saranno utilizzati esclusivamente per scopi accademici (ricerca e didattica dell'interpretazione), escludendo qualsiasi tipo di impiego commerciale.

Fiducioso in una reciproca collaborazione, ringrazio tutti coloro che contribuiranno con la propria disponibilità alla realizzazione del progetto *DIR-SI Corpus*.

Claudio Bendazzoli

Contatti:
EMAIL:
TELEFONO:
FAX:
CELLULARE:

4.2.2.3 Registrazione

Le registrazioni dei materiali inclusi nel corpus, così come di tutti gli altri convegni facenti parte dell'archivio DIRSI-MA, sono state realizzate in prevalenza a cura del *practisearcher*, con la collaborazione dei tecnici di sala di volta in volta presenti. In alcune occasioni, gli stessi tecnici erano stati incaricati dagli organizzatori di registrare il *floor* (in formato digitale, ma talora su supporto magnetico con audiocassette). In casi meno frequenti, per tutto lo svolgimento dei lavori era stato predisposto un servizio di videoregistrazione, affidato a una ditta esterna o al tecnico di sala, a seconda del tipo di strumentazione già presente nella sede del convegno (con o senza cabina regia, videocamere preinstallate e apparecchiature per l'acquisizione audio-video). In genere, gli organizzatori si sono resi disponibili a fornire una copia della registrazione dopo che la ditta incaricata avesse prodotto i relativi CD o DVD. Ciononostante, tali promesse talvolta non sono state successivamente mantenute, per cui è buona prassi effettuare sempre e comunque una registrazione, indipendentemente dalle disposizioni degli organizzatori e dalla disponibilità dei tecnici a fornire loro stessi una copia. Nel caso specifico dei tre convegni inclusi nel corpus, si sono verificate le seguenti condizioni:

- CFF4: TP e TA sono stati registrati autonomamente dal *practisearcher*, con la collaborazione dei tecnici di sala e degli interpreti.
- ELSA: Il TP è stato dapprima registrato su audiocassetta dal tecnico di sala. Al termine del convegno, tutte le audiocassette furono consegnate al *practisearcher*, il quale si era impegnato a digitalizzarle e a fornire una copia dei file digitali anche agli organizzatori. Non fu possibile acquisire autonomamente il *floor* in digitale durante lo svolgimento dei lavori perché il tecnico temeva che si sarebbero creati problemi con l'impianto stereofonico in uso; si decise di non insistere, dato che l'atmosfera si presentava già tesa in partenza: l'impianto di cui era equipaggiata la cabina per l'interpretazione simultanea smise di funzionare poco prima dell'inizio dei lavori, con la conseguente necessità di adottare misure di emergenza per risolvere il problema in tempi brevi.
  Il TA è stato registrato autonomamente dal *practisearcher* con la collaborazione delle interpreti, le quali hanno tenuto un miniregistratore digitale all'interno della cabina.
- CFF5: TP e TA sono stati registrati autonomamente dal *practisearcher*, con la collaborazione dei tecnici di sala e degli interpreti.

La Tabella 6.3 più avanti riporta i contesti e le modalità di registrazione riguardanti tutti i convegni inclusi in DIRSI-MA. Nelle prossime due sezioni sono illustrati gli aspetti più significativi con cui ci siamo confrontati in questa operazione cruciale nella raccolta dei dati.

4.2.2.3.1 Strumentazione tecnica

La gestione diretta della registrazione dei TP e dei TA da parte del *practisearcher* ha comportato la messa a punto di una serie di accorgimenti metodologici e pratici di potenziale interesse anche per altri ricercatori.

Avendo a disposizione un computer portatile con un programma adatto all'acquisizione digitale di segali audio, si è cercato sempre di registrare il TP collegando tale computer portatile all'impianto stereofonico utilizzato in sala e gestito dal tecnico in servizio. Al fine di creare la giusta atmosfera di cooperazione (in uno spirito di *problem-solving*), ci siamo muniti di un "kit tecnologico" per rispondere efficacemente a eventuali richieste da parte del tecnico a cui porre, a nostra volta, una richiesta precisa: poter collegare il nostro computer portatile ad una delle uscite audio dell'impianto in dotazione. Il "kit tecnologico" è stato mano a mano arricchito degli articoli indispensabili a soddisfare le nostre necessità, e siamo giunti alla conclusione che dovrebbe per lo meno comprendere quanto segue: un cavo audio con spinotti RCA (standard in ogni tipo di impianto) e mini jack stereo (da collegare al computer portatile per l'acquisizione), un cavo audio mini jack stereo, un adattatore mini jack stereo, una multipresa, una prolunga e una spina tripla. In questo modo, si contribuisce a evitare di far aumentare le "preoccupazioni" del tecnico, creando una sorta di complicità che va a favore del buon esito della raccolta.

Nel caso in cui una registrazione sia effettuata dal tecnico di sala per conto degli organizzatori, questo non deve trarre in inganno il ricercatore. Infatti, non è detto che sia poi mantenuta la promessa di condividere tutto il materiale non appena sarà pronto il DVD dell'evento, così come potrebbero essere effettuati alcuni tagli alla registrazione, perdendo quindi materiali potenzialmente utili. L'eventuale registrazione predisposta dagli organizzatori, in collaborazione con il tecnico di sala o con una ditta esterna, va semplicemente considerata un ottimo backup dei dati raccolti.

Per quanto riguarda il TA, sono state adottate due diverse strategie con altrettanto diverse strumentazioni. In un primo periodo, è stato utilizzato un miniregistratore digitale a batteria (Olympus DS-660, con una capacità di 11 ore di registrazione), collocato direttamente all'interno della cabina. Nonostante si pensi che la presenza del registratore possa influire sulla validità dei dati, nella nostra esperienza abbiamo constatato che anche con un registratore presente e

visibile[6] all'interno della cabina, gli interpreti perdono la consapevolezza di essere registrati dopo pochi istanti dall'inizio del convegno. L'impegno cognitivo nell'effettuare un'interpretazione simultanea è tale per cui la presenza del registratore o il fatto di essere registrati passa rapidamente in secondo piano, tanto è vero che nel nostro studio sono state molte le occasioni in cui la registrazione non veniva fermata durante la pausa tra una sessione e l'altra per dimenticanza (ovviamente questo non deve succedere nell'attivare la registrazione, come purtroppo è talvolta successo). Un'alternativa interessante sarebbe la possibilità di impostare l'acquisizione del TA assieme all'acquisizione del TP, provvedendo cioè a registrare il tutto su doppia pista. Tuttavia, questa configurazione comporta un impegno maggiore da parte del tecnico, sia in termini di strumentazioni (cavi, regolazione uscite audio e così via), sia in termini di tempo.

Al fine di evitare entrambi gli inconvenienti (richieste eccessive al tecnico di sala e perdita di parti della registrazione), abbiamo prediletto la registrazione del TA con un miniregistratore all'interno della cabina. Tuttavia, è anche possibile ricorrere alla seguente opzione, adottata a un certo punto della raccolta dei dati anche nel presente studio. Poiché almeno uno degli interpreti coinvolti aveva sempre il proprio computer portatile in cabina, è stato possibile acquisire l'audio dell'ambiente della cabina usando un software di registrazione installato nel computer dell'interprete, collegando un microfono di piccole dimensioni direttamente a tale computer.[7] Questa metodologia è di facile realizzazione quando l'interprete è un *practisearcher* (disposto quindi a auto-registrarsi), come nella maggior parte dei convegni inclusi nel presente studio. Come è stato precisato in precedenza (§1.3.2.3.1), la registrazione effettuata in cabina garantisce l'acquisizione di tutti gli scambi comunicativi realizzati dagli interpreti, compresi quelli che avvengono a microfono spento o quando l'interprete disattiva il microfono per un istante premendo il tasto "muto" (mentre si schiarisce la voce o chiede assistenza al collega, ecc.). Bisogna ovviamente tenere conto di questo aspetto nel momento in cui si andranno a produrre le trascrizioni e le clip degli eventi comunicativi ratificati da inserire nel corpus.

---

[6] Il miniregistratore da noi utilizzato ha dimensioni estremamente ridotte e può essere posizionato in modo da risultare non visibile all'interno della cabina. Un accorgimento che vale la pena tenere in considerazione è evitare di collocare il registratore in un punto dove gli interpreti maneggeranno i documenti cartacei a loro disposizione (copie di interventi, glossari, ecc.), così da evitare che il rumore dei fogli disturbi la registrazione delle voci.

[7] A tal fine, è stato usato un microfono standard applicato a un comune set di cuffie. La qualità di registrazione con questo sistema è nettamente superiore a quella offerta dal miniregistratore digitale. Quest'ultimo può essere comunque utilizzato per una registrazione backup di sicurezza, oppure può non essere utilizzato, evitando così di consumare un quantitativo considerevole di batterie.

Riassumendo, la registrazione dei materiali analizzati nel presente studio è disponibile solamente nel formato audio. Il TP è stato acquisito collegando un computer portatile all'impianto di amplificazione delle sale in cui si sono svolti gli eventi, tranne in due casi in cui sono state digitalizzate le audiocassette fornite dagli organizzatori (nel convegno ELSA il tecnico non consentì di collegare il nostro computer all'impianto di amplificazione per timore che questo avrebbe creato disguidi; nel convegno BIRD, la prima sessione è stata recuperata grazie alla registrazione effettuata anche con audiocassette, in quanto il computer portatile utilizzato per l'acquisizione audio si spense a un certo punto della registrazione per non aver impostato correttamente la funzione stand-by e di spegnimento automatico del computer). Dall'altra parte, il TA per un primo periodo di tempo è stato registrato con un registratore digitale collocato all'interno della cabina; successivamente, è stato ottenuto attraverso l'acquisizione audio in digitale, usando il computer portatile all'interno della cabina di uno dei due interpreti.

4.2.2.3.2 Formati e applicazioni per la raccolta e la gestione dei dati

Sulla base della strumentazione usata per effettuare le registrazioni (descritta nella sezione precedente a questa) e al fine di preservare un alto livello di qualità dei dati audio, tutti i materiali sono stati salvati nel formato .WAV. Questo formato è garanzia di alta qualità e compatibilità con la stragrande maggioranza dei programmi di riproduzione. L'unico aspetto sconveniente è che le registrazioni salvate in questo formato risultano "pesanti", hanno cioè dimensioni notevolmente maggiori rispetto a quanto può essere ottenuto con formati compressi come il formato .MP3.

I programmi informatici utilizzati per le acquisizioni e l'editing dei materiali audio sono Cooledit Pro 2.0 e Audacity (versione 1.2.4). Il secondo è un programma scaricabile gratuitamente da internet, ma con funzionalità inferiori rispetto al primo. Le seguenti impostazioni sono state usate per l'acquisizione della traccia audio con i due programmi citati: *sample rate* = 32.000 Hz; *channel* = mono; *resolution* = 16 bit.

Alla pagina successiva, nella Tabella 4.4 sono riassunti gli aspetti più significativi che hanno caratterizzato la procedura di registrazione nei diversi ambienti in cui si sono svolti i convegni raccolti in DIRSI-MA. Nella tabella sono indicati il codice di riferimento del convegno, la durata totale delle registrazioni ottenute, la sede dell'evento, la posizione della cabina degli interpreti e la tipologia, le modalità di acquisizione del segnale audio di TP e TA, nonché gli interpreti effettivamente in servizio e che hanno acconsentito alla registrazione. Anche in questo caso, sono state evidenziate le celle dei convegni inclusi in DIRSI-C.

Tabella 4.4 Caratteristiche ambientali dei convegni e modalità tecniche nella raccolta dei dati DIRSI.

Rif.	Durata registr.	Sede evento	Cabina interpreti	Modalità acquisizione TP	Modalità acquisizione TA	Interpreti in servizio
PTE	**282'**	Sala Salieri presso il Centro Congressi dell'Ente Fiera di Verona (150 posti)	Fissa, in fondo alla sala, frontale ma molto lontano dallo schermo	Digitale, da impianto sala a computer portatile collegato all'uscita audio di un'altra cabina vuota	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-01 (IT-no consenso)
HIST	240'	Sala dei Poeti presso il Dipartimento di Politica, Istituzioni, Storia dell'Università di Bologna (120 posti)	Fissa, in spazio attiguo alla sala in cui si sono svolti effettivamente i lavori (monitor con ripresa fissa b/n disponibile in cabina). Lo spazio in cui è collocata la cabina è un punto di passaggio per le persone che frequentano il Dipartimento	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-01 IT-05
CFF4	**310'**	Sala Convegni presso il Centro Convegni Marani, Ospedale Maggiore, Verona (250 posti)	Fissa, a metà del lato destro della sala guardando il podio. Solo un interprete riesce a vedere (male) lo schermo. Attigui alla cabina vi sono un'altra cabina usata come magazzino (sx) e i servizi igienici per il pubblico (dx)	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto (durante tutto il primo intervento la qualità dell'audio in cuffia e in sala era pessima. Dopo la prima pausa il tecnico è riuscito a sistemare l'impianto, ripristinando una qualità buona del suono)	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-01 UK-01
BIRD	**430'**	Salone polifunzionale senza podio e con sedie removibili presso la sede dell'associazione "Mauro Baschirotto Institute for Rare Diseases" (VI) (100 posti circa)	Mobile, in fondo alla sala, non isolata acusticamente (con una tenda di stoffa al posto della porta) e molto lontana dallo schermo	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto. Solo per la prima parte del convegno, il TP è in realtà stato acquisito in modalità analogica, su supporto magnetico (musicassette), con una successiva digitalizzazione delle parti che per problemi tecnici erano state perse nel corso dell'acquisizione digitale	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-01 IT-02 UK-01
ML10	60'	Sala Convegni presso la sede dell'ICSIM a Villalago (TR) (100 posti)	Fissa, sul lato sinistro della sala guardando il podio, molto vicina allo schermo e al tavolo dei relatori	Registrazione video con videocamera professionale effettuata dagli organizzatori. Successiva acquisizione digitale del solo audio con software apposito da PC	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-01
TICCIH	360'	Sala Convegni (Auditorium Gazzoli) presso Palazzo Gazzoli (TR) (330 posti)	L'impianto di ricetrasmissione è stato predisposto in un camerino attiguo alla sala (divenuto in questo modo la cabina degli interpreti), con un monitor che trasmetteva la ripresa video in diretta, effettuata da un operatore incaricato di videoregistrare l'interno convegno (durante le relazioni spesso la telecamera inquadrava solo l'oratore, impedendo così agli interpreti di vedere anche le diapositive proiettate sullo schermo).	Registrazione video con videocamera professionale effettuata dagli organizzatori (in attesa di essere ricevuta per includerla nell'archivio)	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-01 IT-03
TICCIH	90'	(vedi sopra)	(vedi sopra)	(vedi sopra)	(vedi sopra)	IT-01 IT-03
ELSA	**150'**	Sala polifunzionale presso il Palazzo del Capitano a Cesena (100 posti circa)	Mobile, in fondo alla sala, con una buona visibilità dello schermo, ma con un livello mediocre di isolamento acustico (cabina costituita da pannelli in legno vivo e pannello trasparente in plexiglas)	Analogico, da impianto sala a musicassetta, successivamente digitalizzata con software apposito in PC	Digitale, all'interno della cabina, da miniregistratore a batterie	IT-03 IT-04

Rif.	Durata registr.	Sede evento	Cabina interpreti	Modalità acquisizione TP	Modalità acquisizione TA	Interpreti in servizio
DAY SG	**838'** (tre giorni)	Aula S. Domenico presso l'Ospedale SS: Giovanni e Paolo (VE) (120 posti circa)	Mobile, in fondo alla sala e a una distanza non eccessiva dallo schermo (per un breve momento, i raggi del sole hanno colpito lo schermo entrando da un lucernario posto in alto e non copribile)	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto	Digitale, all'interno della cabina, da computer portatile con microfono esterno standard (lo stesso portatile usato degli interpreti per visualizzare le presentazioni power point autonomamente)	IT-01 UK-01
EDLE SI	180'	Sala Convegni presso l'Archivio di Stato a Terni (60 posti circa)	Mobile, in fondo alla sala e a una distanza non eccessiva dallo schermo	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto	Digitale, all'interno della cabina, da computer portatile con microfono esterno standard (lo stesso portatile usato degli interpreti per visualizzare le presentazioni power point autonomamente)	IT-01 IT-03
CFF5	**307'**	Aula Incontri presso il Centro Convegni Marani, Ospedale Maggiore, Verona (110 posti)	Mobile, in fondo alla sala, frontalmente al podio, con una buona visibilità dello schermo	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto	Digitale, all'interno della cabina, da computer portatile con microfono esterno standard (lo stesso portatile usato degli interpreti per visualizzare le presentazioni power point autonomamente)	IT-01 IT-02
STEE LT	344'	Sala Convegni presso il Centro Multimediale di Terni (136 posti)	Fissa, dentro la cabina regia in fondo alla sala e sopraelevata rispetto al podio; isolata acusticamente dall'ambiente del *floor*, ma non isolata all'interno della cabina regia (con una tenda di stoffa al posto della porta)	Digitale (video) realizzata dagli organizzatori con la collaborazione del tecnico in sala regia	Digitale, all'interno della cabina, da computer portatile con microfono esterno standard (lo stesso portatile usato degli interpreti per visualizzare le presentazioni power point autonomamente)	IT-01 IT-03
CFF7	370'	Aula Incontri presso il Centro Convegni Marani, Ospedale Maggiore, Verona (110 posti)	La stessa di CFF7: mobile, in fondo alla sala, frontalmente al podio, con una buona visibilità dello schermo	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto	Digitale, all'interno della cabina, da computer portatile con microfono esterno standard (lo stesso portatile usato degli interpreti per visualizzare le presentazioni power point autonomamente)	IT-01 UK-01
CFC ARE	320'	Sala convegni presso il Centro Convegni Marani, Ospedale Maggiore, Verona (250 posti)	La stessa di CFF4: fissa, a metà del lato destro della sala guardando il podio. Solo un interprete riesce a vedere (male) lo schermo. Attigui alla cabina vi sono un'altra cabina usata come magazzino (sx) e i servizi igienici per il pubblico (dx)	Digitale, da impianto sala a computer portatile collegato al mixer dello stesso impianto	Digitale, all'interno della cabina, da computer portatile con microfono esterno standard (lo stesso portatile usato degli interpreti per visualizzare le presentazioni power point autonomamente)	IT-01 UK-01

4.2.3 Trascrizione

Le principali questioni teoriche e pratiche relative al faticoso compito di trascrivere i dati orali e al significato da attribuire a tale strumento di analisi sono state approfondite precedentemente (§1.3.3). Come è stato già segnalato, i materiali selezionati per essere inclusi in DIRSI-C sono stati trascritti seguendo le medesime convenzioni adottate nel trascrivere i materiali contenuti nel corpus EPIC. Si tratta di un sistema di trascrizione minimo, con il quale ottenere una base di testo scritto "pulito", a cui poter aggiungere ulteriori livelli di annotazione in futuro. Un altro principio fondamentale a cui si ispira questo sistema di trascrizione è l'equilibrio tra due diversi formati di trascrizione, uno *user-friendly* (cioè leggibile dall'analista secondo metodi tradizionali) e uno *machine-readable* (cioè leggibile dalla macchina, ossia elaborabile al computer). A questo si aggiunge un altro elemento di (auspicato) equilibrio: le trascrizioni dovrebbero essere anche *annotator-friendly*, cioè non dovrebbero richiedere uno sforzo eccessivo al trascrittore per essere prodotte. Anche in questo caso, è stato utilizzato un programma di riconoscimento vocale (Dragon Naturally Speaking), con il quale è stato possibile velocizzare la trascrizione attraverso la tecnica dello *shadowing*. Questa tecnica consiste nell'ascoltare una registrazione e ripeterla simultaneamente a voce alta durante l'ascolto. Regolando la riproduzione delle registrazioni in formato digitale dalla tastiera del computer utilizzato allo scopo di trascrivere i dati, è stato possibile effettuare una vera e propria dettatura al programma di riconoscimento vocale, ottenendo fin dalla prima stesura un risultato decisamente migliore rispetto a quanto si otterrebbe dall'esecuzione dello *shadowing* senza interruzioni di sorta. Inoltre, con l'aumentare delle trascrizioni completate attraverso la dettatura al computer, è andato aumentando anche il tasso di riconoscimento del software.[1]

4.2.3.1 Componente linguistica

Per quanto riguarda la componente linguistica (o verbale), la trascrizione svolta è di tipo ortografico e letterale. Similmente a quanto stabilito per EPIC, in essa sono esplicitati in forma di parola e per esteso i numeri, le date e le cifre; solo nei casi in cui una parola sia emessa secondo una pronuncia non standard, essa è prima normalizzata e poi trascritta letteralmente, così come è stata emessa, inserendola tra due parentesi uncinate e barre, come nel seguente esempio (DIRSI-2006-05-20-VR-CFF4-007b-int-EN-it):

[1] Le trascrizioni del convegno CFF4 sono state realizzate in collaborazione con Marco De Martino, il quale le ha utilizzate per la sua tesi di laurea (2006/2007).

```
ciononostante ha continuato a perdere peso e vedete anche
la flessione della funzionalità </funzionità/> polmonare
```

In genere, dal contesto è possibile risalire alla parola che il soggetto intendeva produrre, cioè "funzionalità" anziché "funzionità" nell'esempio che è stato proposto. Per questo, tale parola è stata prima trascritta secondo la grafia standard, seguita dalla sua versione effettivamente prodotta. Questo accorgimento è indispensabile se si vuole utilizzare un programma per l'annotazione grammaticale automatica. Infatti, qualsiasi *tagger* non sarebbe in grado di attribuire un'etichetta della parte del discorso (*POS-tag*) a eventuali *token* non standard, perché certamente assenti dal suo vocabolario interno (§1.3.4).

4.2.3.2 Componente paralinguistica

Per quanto riguarda la componente paralinguistica (non verbale), sono stati annotati i fenomeni di troncamento, cioè tutte quelle parole che non sono pronunciate completamente o la cui produzione presenta delle "fratture" interne; anche nei casi di troncamento interno, le parole interessate sono prima "normalizzate", cioè trascritte secondo la grafia corretta, seguite dalla forma che riflette la versione orale (nel punto di "frattura" è inserito un trattino basso; si veda l'esempio riportato nella Tabella 6.4).

Nel caso di DIRSI-C, la componente paralinguistica ha subito una semplificazione rispetto alle convenzioni di EPIC, in quanto abbiamo evitato di segnalare la presenza delle pause (sia piene, sia vuote). Questa scelta è stata motivata dalla necessità mantenere una qualità di rappresentazione del parlato ottimale con le scarse risorse umane a disposizione e la tempistica del progetto. Ciò non toglie la possibilità di aggiungere le pause in futuro, attraverso l'uso di strumentazioni informatiche che ne consentano l'individuazione (e quindi l'annotazione) sistematica. Oltre a questo, l'annotazione delle pause e delle vocalizzazioni comporterebbe l'inserimento di almeno un elemento tra un *token* e l'altro (seguendo le convenzioni EPIC, le pause andrebbero annotate scrivendo "..." oppure "ehm" a seconda che siano vuote o piene rispettivamente, Tabella 3.3), e la presenza di tale elemento potrebbe interferire nella sequenza lessicale, ad esempio falsando i risultati di una ricerca sulle collocazioni. Pertanto, tale annotazione dovrebbe essere eseguita in modo da garantire comunque la possibilità di non considerare le pause nel caso di ricerche a livello lessicale.

Le trascrizioni presentano ad ogni modo una segmentazione in unità di significato, i cui confini sono segnalati dall'annotazione di una doppia barra (//) all'interno del testo. Data l'indeterminatezza intrinseca nella definizione di detta

unità (tra l'enunciato e l'unità di informazione), ci siamo basati sulla segmentazione del parlato in unità di analisi derivandole da una sintesi delle indicazioni operative formulate dai diversi autori considerati (§1.3.4.2), con particolare riferimento a Sornicola (1981): la doppia barra che segnala una segmentazione del testo trascritto è posta non solo in base alle indicazioni prosodiche, intonative e semantiche ricavabili dai dati audioregistrati, ma anche in base alla leggibilità della trascrizione da parte dell'analista. In fase di ascolto, costui potrà orientarsi meglio seguendo il flusso di *token* riprodotti sulla carta o sullo schermo, e "raggruppati" in unità di significato.

4.2.3.3 Componente extralinguistica: header

Infine, per la componente extralinguistica è stato sviluppato un apposito *header*, ossia un'intestazione che correda ogni singola trascrizione, apportando informazioni sulla situazione comunicativa, l'evento linguistico e chi l'ha prodotto.[2] I parametri inclusi in questo *header* sono tratti dai parametri brevemente illustrati all'inizio di questo capitolo e corrispondenti alle caratteristiche fondamentali del nostro oggetto di studio. Tuttavia, è stato necessario operare un'integrazione e una rielaborazione degli stessi parametri, in modo da ottenere un'architettura flessibile e rispondere meglio ai tanti casi "ibridi" per i quali sarebbe difficile individuare la funzione dominante di un certo evento linguistico. Alla pagina successiva è riportato lo schema dei parametri strutturati all'interno dell'*header* con le diverse classi di attributi (Tabella 4.5):

[2] I dati contenuti nell'*header* rappresentano già un primo livello di annotazione, in quanto sono rese esplicite diverse informazioni che possono successivamente fungere da "filtro" nell'analisi dei dati.

Tabella 4.5 Parametri inclusi nell'*header* delle trascrizioni DIRSI.

conference title:		(full name)	
conference reference:		CFF4 / CFF5 / ELSA	
conference main topic:		health	
conference date:		year-month-day	
conference location:		Verona / Cesena	
conference session:		opening presentation discussion closing	
session title:		(see official programme)	
speech event:		opening-closing remarks paper or lecture floor allocation procedure or housekeeping announcements question answer comment	
speech number:		000	
speech type:		org-it / org-en / int-it-en / int-en-it	
speech title:		(see official programme and check power point presentation)	
duration:	**timing:**	short medium long	< 900 900-1800 > 1800
speech length:	**number of words:**	short medium long	< 1650 1650-3300 > 3300
speed:	**words per minute:**	low medium high	< 100 100-120 > 120
speech delivery:		impromptu read mixed	
audio visual support:		yes / no	
conference participant:		organizer sponsor chair discussant presenter or lecturer audience interpreter	
conference participant ID:		Surname, Name IT-01 IT-02 IT-03 IT-04 UK-01	
gender:		M / F	
country:		(specify country of origin)	
language:		it / en	
native speaker:		yes / no	
directionality:		A / B	
materials provided to interpreters:		in advance on the spot none	
audio link		(full name of relevant audio file)	
comments:			

Vale la pena soffermarsi su alcuni dei parametri inclusi nell'*header* e sopra riportati assieme ai possibili valori da attribuire a ciascuno di essi. In particolare, meritano un approfondimento i parametri che hanno subito una rielaborazione rispetto alla sintesi precedente (Tabella 4.2), così come i parametri aggiuntivi e i valori soglia attribuibili ad alcuni di essi.

Riguardo ai casi di rielaborazione, si può notare che alcune voci dell'*header* risultano accoppiate, come nei seguenti esempi evidenziati:

speech event:	**opening-closing remarks** **paper or lecture** floor allocation **procedure or housekeeping announcements** question answer comment

conference participant:	organizer sponsor chair discussant **presenter or lecturer** audience interpreter

Questo tipo di abbinamento è stato necessario per meglio rispondere ai tanti casi di eventi linguistici che chiudevano e aprivano una sessione senza soluzione di continuità. Sarebbe stato difficoltoso a livello metodologico (e pratico) spezzare tali eventi linguistici in due sottoeventi, in modo da circoscrivere le due funzioni dominanti in gioco. Oltretutto, avrebbe significato scendere a un livello di annotazione (individuando gli atti linguistici o gli atti comunicativi) che andava ben oltre la portata del presente lavoro. Un esempio di evento linguistico che nell'insieme chiude e apre due diverse sessioni (presentazione e discussione) è il seguente (DIRSI-2007-05-11-VR-CFF5-050-org-it):

```
molte molte grazie e molte complimenti a Gino Galietta per
questa presentazione e per tanti aspetti // sia perché ha
sintetizzato una grossa parte di lavoro sperimentale del
suo laboratorio e collaborativo e di altri laboratori in
maniera molto sintetica // sia perché ha ha semplificato
quanto più possibile un aspetto estremamente specialistico
// io tutte le volte che vado a trovare Gino a Genova vedo
degli strumenti di elettrofisiologia sempre più complicati
// ho sempre più paura di prendere la scossa // e con
questa estrema complessità renderla semplice è veramente
una cosa che solamente quelli bravi bravi riescono a fare
// quindi Gino ci ha fatto un quadro a- abbastanza
panoramico di una serie di approcci farmacologici e anche
ci ha sostanzialmente dato delle informazioni molto molto
di di frontiera su questo campo // e quindi io volevo
sapere se c'era qualcuno che voleva iniziare a rompere il
ghiaccio nel discutere nel discutere quanto ha presentato
Gino in questo questo campo // prego
```

Diversa è la situazione nel seguente esempio, dove la funzione di chiusura di sessione è facilmente determinabile (DIRSI-2006-05-20-VR-CFF4-093-org-it):

```
ecco mi fanno segni che il nostro tempo sarebbe terminato
// quindi se a meno che non ci siano delle domande
irrinunciabili chiuderemmo qua la la sessione </sezione/>
// c'è un rinfresco al piano di sopra // bisogna ritornare
qua all'una e mezza // giusto // ringraziamo molto la
professoressa Moran // molto utile
```

È stato fatto riferimento anche ad alcune integrazioni presenti nello schema di *header* riportato nella Tabella 4.5. Queste riguardano soprattutto elementi contestuali, quali il titolo del convegno, la sigla che abbiamo assegnato a ciascun convegno, il titolo del convegno stesso ed eventualmente delle sessioni e degli eventi linguistici (nel caso si tratti di un tipo di relazione), la data, l'argomento generale, il numero progressivo attribuito a ciascun evento linguistico, il tipo di evento linguistico (distinguendo tra originale e interpretazione, nonché tra lingua e direzione linguistica), il genere e il paese di provenienza dei partecipanti, il nome della clip audio corrispondente (ai fini dell'allineamento testo-suono, §1.3.5.1 e a seguire §4.2.5.1) e uno spazio per eventuali commenti che l'analista può aggiungere in fase di trascrizione.

Infine, restano da esaminare i parametri ai quali sono stati attribuiti particolari valori soglia per differenziare distinte categorie al loro interno. Tra questi, spiccano i valori soglia relativi alla lunghezza, alla durata e alla velocità di emissione degli eventi linguistici, valori che sono stati stabiliti per raggruppare i dati in sottocategorie (per esempio velocità alta, media o bassa)

sulla base dei valori discreti che possono essere calcolati per ciascuno di essi (nel caso della velocità, è stato considerato il numero effettivo di parole al minuto). Dall'altra parte, la definizione di diverse modalità di emissione del TP, distinte nelle tre sottocategorie "impromptu", "read" e "mixed", è basata sull'analisi percettiva di quanto è realmente avvenuto nel contesto di emissione. Anche in questo caso, ci siamo ispirati al modello offerto da EPIC. Tuttavia, considerando le caratteristiche rilevate nei materiali DIRSI (un contesto diverso, il convegno, dalle sedute plenarie del Parlamento), è stato necessario modificare i valori soglia adatti a classificare i materiali EPIC, in modo da riflettere l'andamento delle dinamiche comunicative che si hanno in DIRSI. Nelle seguenti tabelle sono messi a confronto i valori calcolati per questa classe di attributi nei due corpora.

Tabella 4.6 Valori soglia per le sottocategorie di durata (in secondi) degli eventi linguistici in EPIC e DIRSI-C.

Parametro		EPIC	DIRSI-C
Durata (secondi)	short	< 120	< 900
	medium	120-360	900-1800
	long	> 360	> 1800

Tabella 4.7 Valori soglia per le sottocategorie di lunghezza (numero di parole) degli eventi linguistici in EPIC e DIRSI-C.

Parametro		EPIC	DIRSI-C
Lunghezza (numero di parole)	short	< 300	< 1650
	medium	300-1000	1650-3300
	long	> 1000	> 3300

Tabella 4.8 Valori soglia per le sottocategorie di velocità (parole al minuto) degli eventi linguistici in EPIC e DIRSI-C.

Parametro		EPIC	DIRSI-C
Velocità (parole al minuto)	low	<130	< 100
	medium	130-160	100-120
	high	> 160	> 120

4.2.3.4 Sintesi della procedura di trascrizione DIRSI

In termini pratici, per trascrivere i dati registrati è stata seguita questa procedura:

1. Trascrizione integrale del *floor* di ogni singola sessione (o di un interno convegno, a seconda della quantità di dati raccolti) seguendo le convenzioni sopra indicate. Il documento così ottenuto rappresenta una sorta di verbale dell'intero convegno.
2. Numerazione progressiva degli eventi linguistici, partendo da 001 e proseguendo nella numerazione anche nel caso in cui il convegno sia distribuito in diverse giornate (nel senso che la numerazione non è stata ripresa da 001 nelle trascrizioni della seconda giornata).
3. Subito sotto alla numerazione e prima del testo trascritto vero e proprio, annotazione del nome del partecipante, del suo ruolo comunicativo e della categoria di evento linguistico per facilitare l'individuazione dei diversi eventi linguistici.
4. Progressiva compilazione del foglio di lavoro Excel (Tabella 6.1) impostato con una parte dei parametri di classificazione, e utile a monitorare l'andamento del lavoro di trascrizione e di editing dei file audio (produzione delle clip individuali dalle registrazioni dell'intera giornata, sessione o dei gruppi di sessione).
5. Selezione degli eventi linguistici da includere nel corpus, con aggiunta dell'*header* e revisione del testo trascritto.
6. Trascrizione (completa di *header*) del corrispondente TA.

Risulta evidente che i diversi eventi linguistici sono stati "estratti" dall'intero flusso comunicativo a cui appartengono (cominciando dal *floor* e risalendo poi al flusso risultante dalla mediazione dell'interprete) per essere trattati come "testi" individuali e inseriti nella struttura del corpus presentata all'inizio di questo capitolo (§4.2.1). Questo approccio è stato possibile grazie alla particolare microstruttura delle sessioni selezionate per il corpus elettronico, nelle quali abbiamo constatato che il flusso comunicativo risponde a una conformazione più monologica che dialogica, e comunque con un livello di interattività esplicita decisamente inferiore rispetto a quanto avviene nelle sessioni di dibattito o nelle tavole rotonde.
Tutte le trascrizioni sono state prodotte e salvate nel formato .TXT, in modo da ottenere un testo "puro" e leggibile senza problemi di compatibilità da più applicazioni. Il programma TextPad è stato utilizzato non solo in fase di trascrizione e di calcolo del numero di parole,[3] ma anche per la gestione delle trascrizioni stesse, per la correzione dei vari livelli di annotazione e per eseguire

[3] Si tenga presente che il numero di parole così ottenuto non corrisponde al numero di token (§1.3.4.1).

alcune ricerche automatiche, sfruttando una delle funzioni di cui è corredato (cioè la ricerca di particolari stringhe all'interno di uno o più documenti nello stesso formato e raccolti all'interno della medesima directory – una sorta di estrazione automatica di occorrenze, §1.3.2.3.2).

4.2.4 Codifica e annotazione

Una parte di questa tappa nella creazione di DIRSI-C è già stata presentata nella sezione precedente, all'illustrare l'*header* sviluppato per la gestione della componente extralinguistica nelle trascrizioni. Come già anticipato, i diversi parametri di cui si compone l'*header* e i relativi attributi concorrono a formare un nucleo di informazioni, esplicitate all'inizio di ciascuna trascrizione (cioè per ciascun evento linguistico). Successivamente, tali parametri e informazioni così strutturate si possono utilizzare per "filtrare" i dati e svolgere ricerche mirate: per esempio, si potrebbero selezionare i TP prodotti spontaneamente rispetto a quelli prodotti da una lettura o in una modalità mista tra spontaneità e lettura, cioè semispontanei o semipreparati; oppure ci si potrebbe concentrare sui TA emessi con una certa direzionalità, a seconda che si voglia analizzare la resa dell'interprete verso la sua lingua A o B. Ogni singolo parametro è stato compilato manualmente, sulla base dei dati risultanti dal completamento della trascrizione di ogni sessione e dell'archiviazione completa di ciascun convegno in DIRSI-MA. A questo livello di annotazione, se ne aggiungono altri, applicati secondo una modalità automatica o manuale. La prima modalità è stata impiegata per l'annotazione grammaticale, la lemmatizzazione e per codificare l'annotazione delle disfluenze di pronuncia (§4.2.4.2). Dall'altra parte, la seconda modalità è stata usata per l'annotazione temporale (inserendo i *time-tags* in millisecondi con un programma apposito), per poter poi allineare i testi trascritti alle corrispondenti tracce audio (§4.2.4.1).

4.2.4.1 Annotazione temporale

L'annotazione temporale è stata eseguita con il programma Transana (versione 2.12), un programma in cui è presente la funzione di inserimento delle etichette temporali in millisecondi nel testo trascritto, con le quali "agganciare" la traccia audio alla porzione di testo racchiusa da tali etichette (§4.2.5.1). L'operazione è stata eseguita manualmente per ogni singola trascrizione e per tutta la durata della relativa registrazione audio, prima ancora di svolgere l'annotazione automatica (§4.2.4.2). Riportiamo un esempio in cui sono messe a confronto le

due versioni di uno stesso brano di trascrizione, prima e dopo l'annotazione temporale (le etichette temporali sono state evidenziate):

```
DIRSI-2006-05-20-VR-CFF4-001-org-it

Gabriella può chiudere le porte là in
cima per favore // incominciamo //
buongiorno e benvenuti a tutti a questo
quarto seminario detto di primavera che
è un appuntamento annuale // diciamo
dedicato a a a mettere a fuoco alcuni
temi centrali della fibrosi cistica
temi di ricerca temi di cura di
assistenza e di prospettive sul da
farsi //
```

```
DIRSI-2006-05-20-VR-CFF4-001-org-it

Gabriella può chiudere le porte là in
cima per favore ¤<4999> // incominciamo
// buongiorno e benvenuti a tutti a
questo quarto seminario detto di
primavera che è un appuntamento annuale
¤<20756> // diciamo dedicato a a a
mettere a fuoco alcuni temi centrali
della fibrosi cistica // temi di
ricerca temi di cura di assistenza e di
prospettive sul da farsi ¤<35724> //
```

Anche in questo caso, ogni TP e TA è stato trattato come testo "autonomo" e le indicazioni temporali riportate dai *time-tag* devono essere riferite alla sola traccia audio corrispondente, senza mettere in relazione questo tipo di tempistica del TA con la tempistica del TP. In altre parole, eventuali differenze di tempo non sono da imputare al *décalage*, la cui annotazione comporterebbe l'uso di altri programmi informatici e un'altra impostazione (probabilmente a spartito) dei materiali trascritti (§1.3.3).

Come è già stato puntualizzato nel primo capitolo (Leech 1997a, pp. 6-8; §1.3.4), è essenziale preservare una copia "grezza" delle trascrizioni, vale a dire una versione priva di tutti i successivi livelli avanzati di annotazione; pertanto, il mantenimento di una copia priva dell'annotazione temporale di tutte le trascrizioni è fondamentale per la buona gestione futura di tutti i materiali. Conseguentemente, l'applicazione delle etichette temporali comporta la creazione di una prima copia di tutto il materiale trascritto, "arricchita" dei *time-tag*. Questo potrebbe trasformarsi in una "pericolosa" proliferazione dei materiali trascritti.[4] A questo proposito, l'organizzazione di tutti i materiali in cartelle e sottocartelle deve essere impostata in modo organico e con la massima cura da parte di chi gestisce la ricerca.

[4] L'esistenza di più copie contenenti diversi livelli di annotazione diventa potenzialmente complessa da gestire in fase di correzione delle trascrizioni, poiché ogni eventuale errore dovrebbe essere corretto in tutte le copie e non solo nella versione "grezza".

4.2.4.2 Annotazione grammaticale, lemmatizzazione e codifica delle disfluenze di pronuncia

Il sistema di codifica e annotazione adottato in DIRSI è lo stesso che è stato applicato ai materiali EPIC. Per questo, le trascrizioni sono state elaborate con due programmi in linguaggio Perl creati appositamente per EPIC,[5] e da noi modificati in funzione del nuovo *header* presente in DIRSI (i due programmi sono stati denominati `tagging_dirsi.pl` ed `encoding_dirsi.pl`). Implementando il programma per la codifica e l'annotazione dei file di testo contenenti le trascrizioni (`tagging_dirsi.pl`), si sono ottenuti quattro file (uno per ciascuna lingua e tipologia, cioè originali italiani, originali inglesi, interpretazioni italiane e interpretazioni inglesi). In questo modo, tutti gli eventi linguistici originali italiani risultano strutturati all'interno di in un unico file (DIRSI-ORG-IT), corrispondente al corpus dei TP in italiano; lo stesso vale per gli eventi linguistici originali inglesi, tutti raccolti nello stesso file (DIRSI-ORG-EN) contenente i TP in inglese; similmente, le trascrizioni delle interpretazioni risultano tutte codificate in due diversi documenti, a seconda della direzione linguistica: un documento con i TA in italiano (DIRSI- INT-EN-IT) e un documento con i TA in inglese (DIRSI-INT-IT-EN). Si arriva così alla struttura del corpus di cui abbiamo dato una rappresentazione grafica all'inizio di questo capitolo nella Figura 4.3 e che è riproposta sotto con una diversa veste grafica:

DIRSI-ORG-IT **DIRSI-ORG-EN**

DIRSI-INT-IT-EN **DIRSI-INT-EN-IT**

All'interno di questi documenti,[6] le trascrizioni annotate risultano impostate su quattro diverse colonne, secondo una struttura modulare e compatibile con i programmi della CWB – *Corpus Work Bench* (§1.3.4). Riprendendo lo stesso esempio richiamato prima in merito all'annotazione delle disfluenze di pronuncia, ecco come appare la stessa porzione di trascrizione dopo che è stata

[5] L'autore che ha realizzato i programmi per codificare e indicizzare EPIC è il prof. Marco Baroni dell'Università di Trento. Grazie alla sua preziosa consulenza nel corso della nostra ricerca e alle nozioni apprese durante il soggiorno di studio presso il laboratorio LLI della Universidad Autónoma de Madrid (settembre-dicembre 2007) siamo riusciti a modificare tali programmi per poter codificare, annotare e indicizzare i materiali raccolti in DIRSI-C.

[6] Per facilitare la gestione dei dati in questa tappa della creazione del corpus, la codifica e l'annotazione (compresa la sua revisione) sono state eseguite dapprima su sottocorpora specifici contenenti solo i dati di ogni singolo convegno. In altre parole, sono stati prima creati tre corpora parziali (DIRSI_CFF4, DIRSI_ELSA e DIRSI_CFF5), a loro volta costituiti da quattro sottocorpora ciascuno, come nella struttura generale di DIRSI-C (org-it, int-it en, org-en e int-en-it). Dall'unione dei vari sottocorpora appartenenti ai tre convegni sono stati ottenuti i sottocorpora di cui si compone DIRSI-C (Figura 4.3, §4.2.1).

elaborata dal programma di annotazione e codifica `tagging_dirsi.pl` (si noti il penultimo *token* affetto dalla pronuncia non standard):

ha	ha	avere	VER:pres
continuato	continuato	continuare	VER:pper
a	a	a	PRE
perdere	perdere	perdere	VER:infi
peso	peso	peso	NOM
e	e	e	CON
vedete	vedete	vedere	VER:pres
anche	anche	anche	ADV
la	la	il	DET:def
flessione	flessione	flessione	NOM
della	della	del	PRE:det
funzionalità	/funzionità/	funzionalità	NOM
polmonare	polmonare	polmonare	ADJ

Come si vede nell'esempio sopra riportato, la prima colonna è riservata alla trascrizione tokenizzata, cioè ai *token* trascritti (e normalizzati), incolonnati uno sotto l'altro; la seconda colonna è riservata alla trascrizione letterale, riportante quindi tutti i casi annotati di disfluenze di pronuncia e di troncamento interno racchiusi tra due barre; la terza colonna è riservata al lemma; infine, la quarta colonna è riservata alle etichette dell'annotazione grammaticale, cioè i *pos-tag*. All'inizio di ogni trascrizione, l'*header* risulta invece strutturato con tutti i suoi attributi in formato XML. Nell'immagine sotto (Figura 4.6), è offerta una rappresentazione di uno dei file con la struttura che è stata descritta:

Figura 4.6 Esempio di trascrizione DIRSI codificata, annotata e strutturata secondo un formato modulare compatibile con CWB.

```
<corpus>
<speech conftitle="participation and partnership in the delivery of services supporting elderly people and their carers" confref="ELSA" confmaintopic="health" confdate="2006-10-19" conflocation="Cesena" confsession="opening" sessiontitle="NA" speechevent="opening-closing remarks" speechid="001" speechtype="org-it" speechtitle="NA" duration="short" timing="148" speechlength="short" words="371" speed="high" wordsperminute="150" delivery="impromptu" avsupport="no" participant="chair" participantid="Leonardi, Barbara" gender="F" country="Italy" lang="it" nativespeaker="yes" directionality="NA" materials2interpreters="NA" audiolink="DIRSI-2006-10-19-FC-ELSA-001-org-it" comments="NA">
allora      allora      allora      ADV
ce          ce          ce          PRO:pers
l'          l'          il          DET:def
ab-         ab-         UNKNOWN     TRUNC
ce          ce          ce          PRO:pers
l'          l'          il          DET:def
abbiamo     abbiamo     avere       VER:pres
fatta       fatta       fare        VER:pper
¤<3936>     ¤<3936>     UNKNOWN     TT
//          //          //          SENT
e           e           e           CON
iniziamo    iniziamo    iniziare    VER:pres
con         con         con         PRE
un          un          un          DET:indef
po'         po'         po'         ADV
di          di          di          PRE
ritardo     ritardo     ritardo     NOM
scusateci   scusateci   scusare     VER:impe
¤<7440>     ¤<7440>     UNKNOWN     TT
//          //          //          SENT
```

Il programma di codifica e annotazione (`tagging_dirsi.pl`) attinge a uno dei *tagger* descritti nel primo capitolo (§1.3.4.1), ovvero il Treetagger, sia per l'italiano (versione standard), sia per l'inglese. I repertori di etichette (*tagset*) abbinati alle due versioni linguistiche di Treetagger (§1.3.4.1, Tabella 1.10 e Tabella 1.11) sono stati integrati con l'aggiunta di un *tag* nel *tagset* italiano per annotare la presenza di parole straniere (FW = *foreign word*), in quanto lo stesso *tag* risultava già incluso nel *tagset* inglese; inoltre, a entrambi i tagset sono stati aggiunti il *tag* TRUNC per le parole troncate e il *tag* TT per l'annotazione temporale (§4.2.4).

Tutti i sottocorpora annotati sono stati interamente revisionati e corretti manualmente. Nonostante il buon tasso di successo garantito dall'etichettatore automatico scelto (Sandrelli & Bendazzoli 2006), si è preferito svolgere una revisione globale per via dei diversi casi di annotazione errata o di lemmatizzazione non riuscita che risultavano comunque presenti. Curiosamente, le due versioni di Treetagger (una per l'italiano e una per l'inglese) si sono comportate in modi diversi di fronte a problemi identici di annotazione. Per esempio, si sono verificati diversi casi di parole sconosciute al *tagger*, per le quali è assegnata comunque una particolare etichetta di lemmatizzazione (`UNKNOWN`). Da una parte, il *tagger* inglese ha sempre annotato

grammaticalmente tali occorrenze come sostantivi (`NN`); dall'altra parte, il *tagger* italiano ha comunque cercato di applicare l'etichetta più plausibile nel contesto in cui tale occorrenza si verificava, per cui i casi di lemmatizzazione sconosciuta presentavano diverse categorie di *pos-tag*. Questo ha ovviamente comportato uno sforzo maggiore in fase di correzione. Nello specifico, alcuni casi problematici hanno riguardato termini non presenti nel lessico interno dei *tagger*, quali i termini specialistici (ad esempio *transmembrane*, *pharmacotherapy*, *mucociliary*, *druggable*, ecc.) e i numerali superiori a venti. Il *tagger* italiano ha registrato casi di annotazione errata anche con termini più comuni; un caso emblematico è la discriminazione tra le categorie congiunzione/pronome relativo del *token* "che". Ad esempio, nelle trascrizioni dei TP selezionati dal convegno CFF4, su un totale di 110 occorrenze di "che", solo 6 erano state taggate correttamente. Abbiamo constatato che tendenzialmente il *tagger* italiano attribuisce l'etichetta `PRO:rela` (pronome relativo) a tutte le occorrenze di "che", con rari casi di attribuzione dell'eticheta `CON` (congiunzione).[7]

Al fine di svolgere il compito di revisione e correzione manuale, la già citata funzione di ricerca presente in TextPad è stata uno strumento di fondamentale importanza. Essa ha infatti consentito non solo di individuare velocemente gran parte dei casi critici, ma anche di richiamare il file della trascrizione corrispondente, nonché di svolgere sostituzioni e correzioni automatiche nei casi di errori frequenti e ripetuti.[8]

A completamento della revisione, i testi annotati possono essere indicizzati con il secondo programma da noi adattato (`encoding_dirsi.pl`), per poter così essere analizzabili automaticamente con i programmi inclusi nel pacchetto *IMS Corpus Work Bench* (CWB). Tali programmi consentono di interrogare il corpus attraverso *queries* (richieste di informazione) formulate secondo le regole sintattiche del linguaggio *Corpus Query Processor* (CQP).

Grazie ai formati in cui sono stati predisposti i materiali trascritti, al pari dei materiali codificati e taggati, essi si sono dimostrati compatibili con un altro sistema di indicizzazione, rispondente a un tipo di etichettatura e codifica su base XML (§1.3.4). Questa versatilità si è rivelata fondamentale al fine di poter analizzare il corpus con altri strumenti di ricerca oltre a CWB. Partendo dai documenti codificati e annotati secondo il formato sopra descritto, gli attributi

[7] Nei TP e TA italiani selezionati dal convegno CFF5 si sono ottenuti i seguenti risultati: nei TP italiani, su 492 occorrenze totali di "che", solo 15 risultavano taggate automaticamente come CON, di cui 3 in modo errato. Nei TA in italiano, solo 11 occorrenze su 412 risultavano taggate come CON (di cui due casi sono errati).

[8] Tutti i casi di criticità più rilevanti sono stati raccolti in un documento che ha guidato la revisione globale di tutti i materiali annotati.

posizionali espressi nel formato modulare per ogni singolo *token* e i relativi attributi strutturali inclusi nell'*header* sono stati convertiti in un formato di etichettatura XML; rispetto all'immagine proposta prima (Figura 4.6), questa trasformazione ha predisposto le trascrizioni nel modo di seguito rappresentato (Figura 4.7):

Figura 4.7 Esempio trascrizione DIRSI codificata, annotata e strutturata in XML.

```
<?xml version="1.0" encoding="utf-8"?>
<fml>
<Header>
<Conference title="participation and partnership in the delivery of services supporting elderly people and their carers" reference="ELSA" topic="health" date="2006-10-19" location="Cesena"/>
<Session type="opening" title="NA">
<Speech id="001" type="org-it" event="opening-closing remarks" title="NA" duration="short" timing="148" length="short" words="371" speed="high" wordsperminute="150" delivery="impromptu" avsupport="no" materials2interpreters="NA"/>
<Participant role="chair" name="Leonardi, Barbara" shortname="LEO" sex="F" country="Italy" language="it" nativespeaker="yes" directionality="NA"/>
<Audiolink> DIRSI-2006-10-19-FC-ELSA-001-org-it </Audiolink>
<Comments> NA </Comments>
</Header>
<Text>
<p>
<f h="LEO" st="0" et="3.936" id="1">
        <w tok="allora" trans="allora" lem="allora" pos="ADV" id="1-1-1"> allora </w>
        <w tok="ce" trans="ce" lem="ce" pos="PRO:pers" id="1-1-2"> ce </w>
        <w tok="l'" trans="l'" lem="il" pos="DET:def" id="1-1-3"> l' </w>
        <w tok="ab-" trans="ab-" lem="UNKNOWN" pos="TRUNC" id="1-1-4"> ab- </w>
        <w tok="ce" trans="ce" lem="ce" pos="PRO:pers" id="1-1-5"> ce </w>
        <w tok="l'" trans="l'" lem="il" pos="DET:def" id="1-1-6"> l' </w>
        <w tok="abbiamo" trans="abbiamo" lem="avere" pos="VER:pres" id="1-1-7"> abbiamo </w>
        <w tok="fatta" trans="fatta" lem="fare" pos="VER:pper" id="1-1-8"> fatta </w>
</f>
</p>
<p>
<f h="LEO" st="3.936" et="7.44" id="2">
        <w tok="e" trans="e" lem="e" pos="CON" id="2-1-1"> e </w>
        <w tok="iniziamo" trans="iniziamo" lem="iniziare" pos="VER:pres" id="2-1-2"> iniziamo </w>
        <w tok="con" trans="con" lem="con" pos="PRE" id="2-1-3"> con </w>
        <w tok="un" trans="un" lem="un" pos="DET:indef" id="2-1-4"> un </w>
        <w tok="po'" trans="po'" lem="po'" pos="ADV" id="2-1-5"> po' </w>
        <w tok="di" trans="di" lem="di" pos="PRE" id="2-1-6"> di </w>
        <w tok="ritardo" trans="ritardo" lem="ritardo" pos="NOM" id="2-1-7"> ritardo </w>
        <w tok="scusateci" trans="scusateci" lem="scusare" pos="VER:impe" id="2-1-8"> scusateci </w>
</f>
</p>
```

Questo secondo sistema di codifica è simile a quello sviluppato dai ricercatori del Laboratorio de Lingüística Informática (LLI) presso la Universidad Autónoma de Madrid (UAM), dai quali è stata ottenuta una preziosa collaborazione in questa fase di realizzazione del corpus. La caratteristica più significativa di questo sistema di codifica e annotazione sta nella segmentazione in enunciati o unità di informazione, delimitate dalle etichette temporali. Questo tipo di segmentazione è di ordine superiore alla segmentazione in unità di

significato da noi realizzata con l'inserimento della doppia barra nel corpo della trascrizione. Infatti, se quest'ultima risponde soprattutto alla necessità di rendere la trascrizione immediatamente leggibile a occhio nudo (rispettando le unità di significato, e quindi separando, per esempio, gli elementi di una elencazione che risulterebbe difficoltoso "sentire con gli occhi"), la prima rispecchierebbe più da vicino la scansione effettiva in enunciati e unità di informazione.

4.2.5 Allineamento

Le diverse possibilità di allineamento in un corpus di interpretazione simultanea sono state presentate nel primo capitolo (§1.3.5), con la distinzione generale tra due tipi di allineamento: l'allineamento testo-suono, ossia il collegamento diretto tra una trascrizione e la sua traccia audio/video corrispondente, e l'allineamento tra TP e TA, realizzabile sia a livello di contenuto, sia a livello di *décalage*. Le modalità di allineamento effettivamente applicate a DIRSI-C sono l'allineamento testo-suono per ogni singolo evento linguistico e l'allineamento TP-TA sulla base del contenuto. L'allineamento tra il testo trascritto e la corrispondente traccia audio nei materiali DIRSI (descritto nella sezione successiva §4.2.5.1) è stato realizzato grazie all'annotazione temporale presentata nella sezione precedente (§4.2.4.1). La presenza dei *time-tag* nel corpo delle trascrizioni, e quindi all'interno dei documenti codificati e annotati, ha consentito l'applicazione di un sistema di allineamento già sperimentato in altri progetti di ricerca sugli spoken corpora (Moreno Sandoval et al. 2005). Dall'altra parte, l'allineamento TP-TA (§4.2.5.2) è stato eseguito in maniera semiautomatica e grazie alla collaborazione dei ricercatori del LLI-UAM.

4.2.5.1 Allineamento testo-suono

Al fine di stabilire un collegamento tra le clip audio e le trascrizioni facenti parte di DIRSI-C, sono state eseguite due operazioni preliminari fondamentali: è stato inserito all'interno dell'*header* un parametro dedicato a tale scopo (`audio link`), riservato all'indicazione del nome esatto del file multimediale da collegare al file di testo della trascrizione interessata;[9] è stata effettuata l'annotazione temporale delle singole trascrizioni, inserendo etichette riportanti i millisecondi dei punti di inizio e fine degli enunciati all'interno del corpo della

[9] Seguendo il nostro schema di composizione delle denominazioni dei file, si avrà la stessa denominazione per la clip in formato .WAV e per la trascrizione in formato .TXT (cambia solo il formato).

trascrizione, creando così degli "ancoraggi" con la traccia audio corrispondente. Con i materiali così impostati, si è proceduto all'annotazione automatica delle parti del discorso e del lemma, assieme alla codifica delle disfluenze di pronuncia, dell'*header* e dell'intero "corpus" di trascrizioni suddivise per lingua e per tipologia (originali e interpretazioni). Tutti i file ottenuti al termine di questa procedura sono stati convertiti automaticamente (con un programma apposito in linguaggio Perl, `CONVER_TRANSANA.PL`) nel formato utilizzato per i materiali spagnoli appartenenti al progetto C-ORAL-ROM (Moreno Sandoval et al. 2005). In tale formato, il testo delle trascrizioni è segmentato in enunciati, creati a partire dalle etichette temporali che segnano il punto di inizio e fine degli stessi enunciati. Implementando il programma di conversione, si ottiene una codifica su base XML identica a quella rappresentata precedentemente nella Figura 4.7. Lo stesso sistema è stato applicato anche ai materiali di un altro corpus orale in lingua spagnola, ovvero il corpus CHIEDE (Garrote 2008) sul linguaggio infantile, da cui abbiamo estratto il seguente esempio:

```
<UNIT speaker="TEA" startTime="0" endTime="0.509"> a ver Daniel </UNIT>

<UNIT speaker="TEA" startTime="0.509" endTime="1.559"> ¿ cuántos años tienes tú ?
</UNIT>

<UNIT speaker="DAI" startTime="1.559" endTime="2.26"> cinco </UNIT>
```

L'etichetta `<UNIT>` racchiude un singolo enunciato ed è corredata di attributi che ne specificano l'autore, il tempo di inizio e la fine. Grazie a queste informazioni così esplicitate, è possibile risalire alla porzione corrispondente all'interno del relativo file multimediale, il quale risulta "segmentato" e riproducibile, solo per la porzione interessata, con un programma di riproduzione (*mediaplayer*) opportunamente collegato.

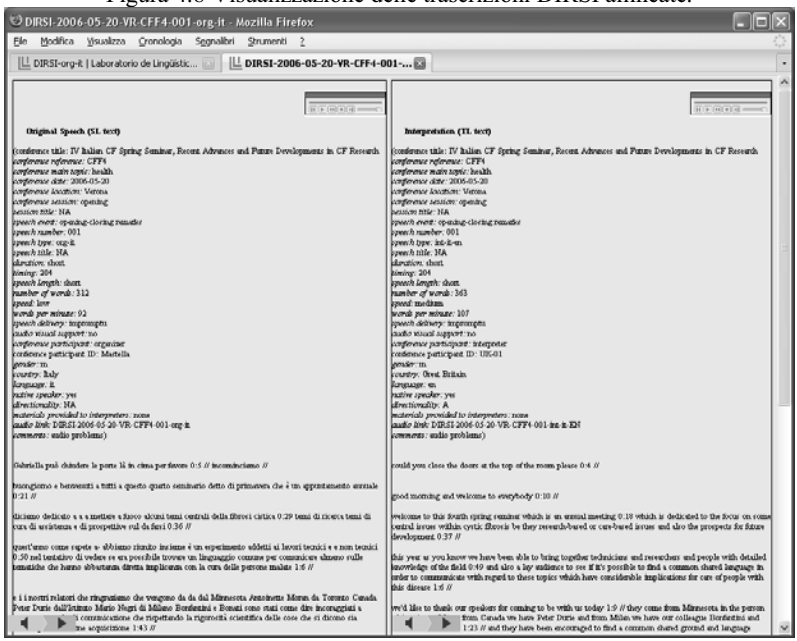
4.2.5.2 Allineamento TP-TA

Quest'altra modalità di allineamento è stata effettuata con una procedura semiautomatica. Ciascuna coppia di eventi linguistici (cioè ogni TP con il corrispondente TA) è stata inserita in una pagina web strutturata su due colonne. Nella colonna di sinistra trova spazio il TP, mentre la colonna di destra è riservata al TA. Implementando un programma creato appositamente, è stato possibile configurare le trascrizioni in questo modo, visualizzando i testi affiancati e mandando a capo la linea di trascrizione ogni qualvolta che fosse presente la doppia barra (inserita per annotare la segmentazione del testo in unità

di significato come descritto sopra). Ovviamente, le diverse condizioni di produzione dei TP e dei TA hanno fatto sì che la segmentazione delle trascrizioni non combaciasse perfettamente; è superfluo specificare che anche lo stesso numero di parole in ciascun brano all'interno di una trascrizione non trova mai una corrispondenza perfetta tra l'originale e l'interpretazione. Per tutti questi motivi, i due testi affiancati e distribuiti automaticamente su diverse righe presentavano sempre uno sfasamento. Un intervento di revisione manuale (utilizzando KompoZer 0.7.10, un editor di pagine web scaricabile gratuitamente da Internet) ha permesso di regolare la distribuzione dei due testi affiancati, di modo che risultassero ben allineati sulla base del contenuto. Sicuramente, si otterrebbe una visualizzazione ben diversa se l'allineamento fosse regolato in base allo sviluppo temporale effettivo di ciascun flusso comunicativo, seguendo cioè il *décalage* dell'interprete rispetto all'emissione del TP. Tuttavia, questo richiederebbe un diverso approccio alla trascrizione e all'allineamento stesso.

Tutte le pagine web con le trascrizioni allineate sono corredate di due riproduttori audio per poter ascoltare la clip corrispondente. Un primo tipo di riproduttore è visualizzato in alto a destra di ciascuna colonna. Questi *mediaplayer* consentono di ascoltare la clip del TP in sovrapposizione alla clip del TA. Pur non essendo questo un metodo scientifico per riprodurre il *décalage* reale tra TP e TA, consente comunque di svolgere un ascolto realistico delle due registrazioni in parallelo. L'altro tipo di riproduttore è collocato nella parte inferiore dello schermo e rimane in tale posizione allo scorrere in basso della pagina. Pertanto, con questo *mediaplayer* ci si può posizionare nel punto desiderato, far partire la registrazione e regolarla al punto esatto grazie alle indicazioni delle etichette temporali che sono visibili ed evidenziate nel corpo della trascrizione stessa. Di nuovo, si deve tenere presente che le indicazioni di tempo presenti nella colonna del TP non sono da mettere in correlazione con le stesse indicazioni presenti nella colonna del TA. Esse mettono in correlazione solo ed esclusivamente un testo trascritto con la sua clip audio corrispondente. L'immagine sotto riportata (Figura 4.8) mostra un esempio di questo tipo di visualizzazione, messa a punto per l'allineamento TP-TA (e testo-suono) di tutte le trascrizioni in DIRSI-C:

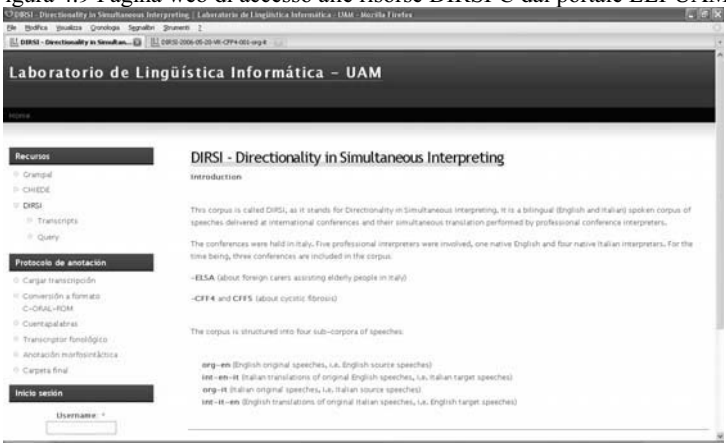
Figura 4.8 Visualizzazione delle trascrizioni DIRSI allineate.



4.2.6 Accessibilità al corpus

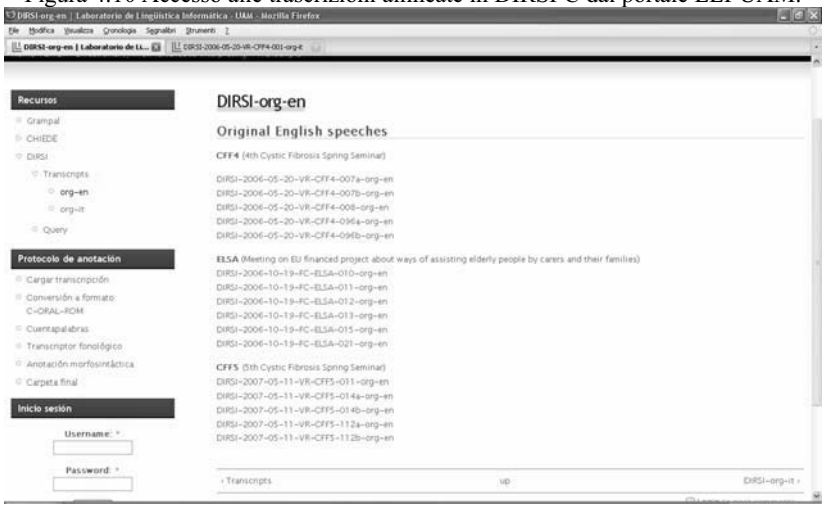
Grazie alla versatilità dei formati in cui sono disponibili i materiali DIRSI, essi possono essere esplorati utilizzando molteplici piattaforme e strumenti di linguistica computazionale. Oltre alla *suite* di programmi contenuti in CWB (utilizzabili per esempio in ambiente UNIX e accedendovi con un *client* SSH come Putty, scaricabile gratuitamente da Internet), i materiali contenuti in DIRSI-C sono accessibili dal portale del Laboratorio LLI-UAM (http://drusila.lllf.uam.es). L'accesso è impostato secondo due differenti modalità: Transcript e Query, come si può vedere nell'immagine sotto riportata (Figura 4.9):

Figura 4.9 Pagina web di accesso alle risorse DIRSI-C dal portale LLI-UAM.



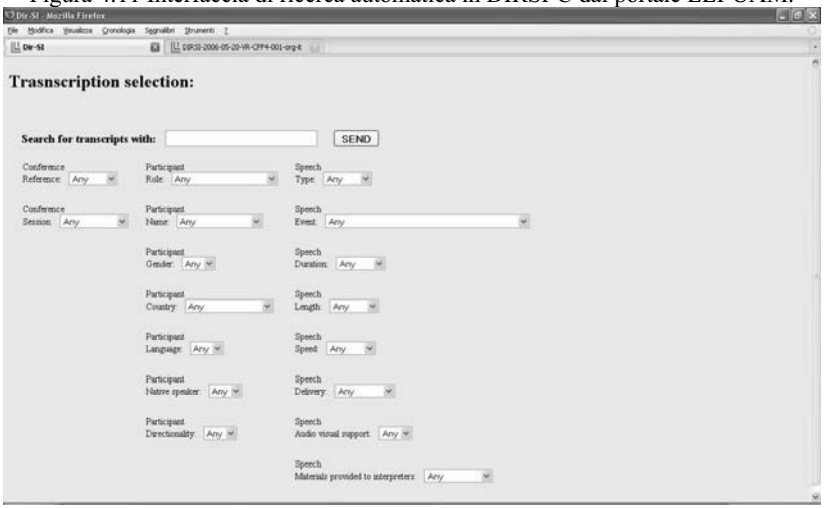
Cliccando sul link Transcript si possono visualizzare le trascrizioni allineate sulla base del contenuto, con incluso il collegamento alla traccia audio corrispondente. Dopo aver scelto se si vuole accedere ai TP italiani (con i relativi TA in inglese) o ai TP inglesi (con i relativi TA in italiano), appare l'elenco di tutti i file di trascrizione, suddivisi per convegno (Figura 4.10). Ogni link dà accesso alla pagina web strutturata su due colonne, contenenti il TP e il TA allineati sulla base del contenuto. Per poter visualizzare correttamente i *mediaplayer* e avere così anche accesso alle tracce audio, è necessario utilizzare il browser FireFox, facendo attenzione ad impostare adeguatamente la dimensione della pagina:

Figura 4.10 Accesso alle trascrizioni allineate in DIRSI-C dal portale LLI-UAM.



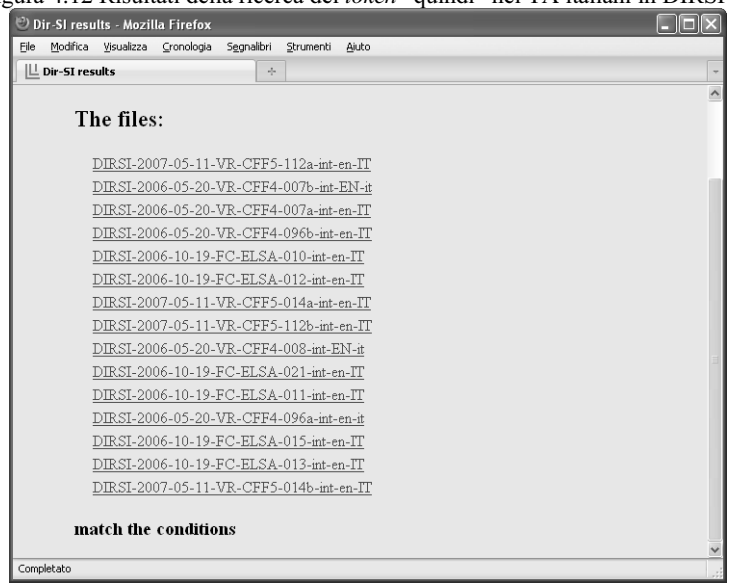
Dall'altra parte, cliccando sul link Query presente nel menù a sinistra, si ha accesso a una interfaccia di ricerca, corredata di filtri con cui poter restringere le ricerche in base a diversi parametri relativi all'evento comunicativo generale (selezione convegni e/o sessioni), ai partecipanti e all'evento linguistico. Le ricerche possono essere effettuate per singoli *token* o per stringhe di *token*, sia come ricerca libera, sia con la specificazione dell'etichetta grammaticale. L'immagine riportata nella Figura 4.11 mostra una schermata dell'interfaccia che è stata sviluppata:

Figura 4.11 Interfaccia di ricerca automatica in DIRSI-C dal portale LLI-UAM.



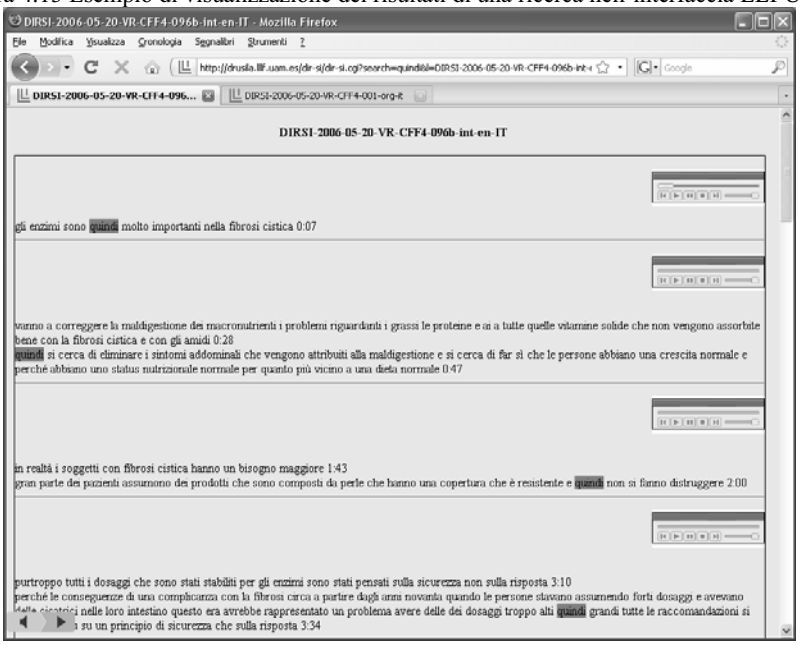
Dopo ogni ricerca, si ottiene una pagina di risultati in cui sono mostrati i link ai brani di trascrizione in cui sono presenti i *token* richiesti. Per fare un esempio, si veda nella seguente immagine (Figura 4.12) la pagina di risultati ottenuta dalla ricerca del *token* `quindi` nei TA in italiano:

Figura 4.12 Risultati della ricerca del *token* "quindi" nei TA italiani in DIRSI-C.



Cliccando su uno dei link contenuti nella pagina dei risultati, si accede alla visualizzazione di tutte le occorrenze presenti nella stessa trascrizione (è mostrato ogni enunciato in cui appare il *token*), con un link alla porzione della clip audio corrispondente, in modo da poter ascoltare il dato reale a cui fare riferimento per affinare l'analisi. Continuando con l'esempio introdotto, abbiamo selezionato il link della trascrizione DIRSI-2006-05-20-VR-CFF4-096b-int-en-IT per visualizzare tutte le occorrenze di `quindi` che sono presenti in tale trascrizione (Figura 4.13):

Figura 4.13 Esempio di visualizzazione dei risultati di una ricerca nell'interfaccia LLI-UAM.



Il *mediaplayer* a destra di ogni brano di trascrizione riproduce l'enunciato visualizzato, mentre il *mediaplayer* in basso è collegato alla traccia audio intera. Da alcune ricerche a campione, abbiamo verificato che non tutti i link sono impostati correttamente. Sarà pertanto necessario effettuare una futura correzione dei collegamenti problematici mano a mano che questi vengono individuati.

Per risalire al TP corrispondente (o al TA nel caso si esegui una ricerca sui soli TP) è sufficiente rifarsi alle informazioni ottenibili dalla denominazione dei singoli file. Si può decidere di visualizzare le trascrizioni allineate, per poi

svolgere una ricerca del *token* corrispondente selezionando il tipo di evento linguistico complementare a quello che si è cercato prima (TP o TA).

Dall'integrazione delle diverse ricerche che possono essere effettuate sia con CWB, sia con l'interfaccia presente nel portale LLI-UAM, si aprono molteplici percorsi di ricerca che si integrano e arricchiscono a vicenda.

4.2.6.1 Condizioni d'uso e di distribuzione

La distribuzione e l'uso dei materiali raccolti in DIRSI-MA e DIRSI-C devono sottostare al rispetto di precise condizioni, in quanto la raccolta stessa dei materiali è avvenuta sulla base di un consenso informato sottoscritto dai partecipanti (§4.2.2.2). Al fine di tutelare la privacy dei soggetti registrati e di non venire meno alle condizioni accordate in fase di raccolta, queste ultime sono state integrate dalle seguenti:

1. L'anonimato degli interpreti deve essere mantenuto sempre.

2. Tutte le persone e le istituzioni menzionate non devono essere contattate per finalità inerenti alle iniziative di ricerca che deriverebbero dall'uso dei materiali DIRSI, né per proporre servizi linguistici o altre attività commerciali.

3. I materiali non devono essere divulgati o distribuiti a terzi. Eventuali richieste devono essere rivolte al responsabile del progetto.

4. L'uso dei materiali DIRSI implica il rispetto delle condizioni indicate nel modello di consenso informato, utilizzato per la registrazione dei materiali (in particolare, uso esclusivo per fini accademici – ricerca e didattica).

4.3 Descrizione di DIRSI-C

I dati effettivamente inclusi all'interno del corpus elettronico DIRSI-C sono una selezione rappresentativa dell'intera campionatura dei dati raccolti e immagazzinati nell'archivio multimediale DIRSI-MA. In questa sezione presenteremo i dati del corpus e ne illustreremo le principali caratteristiche, mettendole a fuoco grazie all'estrazione automatica delle informazioni più rilevanti tra quelle racchiuse nell'*header* e quelle ricavabili da alcuni dei livelli di annotazione applicati alle trascrizioni. Per ogni caratteristica saranno forniti non solo i dati globali riferiti all'intero corpus, ma anche i dati specifici riferiti ai singoli convegni (CFF4, ELSA e CFF5). Come è già stato puntualizzato, ciascun convegno potrebbe essere considerato un microcosmo a se stante, e quindi un piccolo corpus composto da quattro sottocorpora (org-it, int-it-en, org-en e int-en-it), come lo è DIRSI-C con la sua struttura globale (Figura 4.3). L'osservazione delle tendenze ottenute su piccola scala è utile al fine di individuare eventuali differenze particolari tra i singoli eventi comunicativi mediati.

Prima di descrivere i dati presenti nel corpus in riferimento ai diversi parametri di classificazione che sono stati adottati, l'aspetto di maggior interesse concerne sicuramente la dimensione globale del corpus. La Tabella 4.9 riporta il numero di parole totale (oltre 135.000) e il numero di parole per ciascun sottocorpus, con il relativo numero di testi (ovvero di eventi linguistici):

Tabella 4.9 Dimensione (numero di parole) totale di DIRSI-C.

sottocorpus	numero di testi	numero di parole	% di DIRSI-C
ORG-IT	63	33.412	24,6
INT-IT-EN	63	31.510	23,2
ORG-EN	16	37.249	27,4
INT-EN-IT	16	33.664	24,8
TOTALE	158	135.835	100

Come si può desumere dai dati percentuali espressi nella colonna a destra, DIRSI-C è un corpus piuttosto bilanciato, con una lieve sovra-rappresentazione dei TP inglesi (per il numero di parole). Un'altra osservazione generale è che i

sottocorpora di TA contengono sempre meno parole dei rispettivi sottocorpora di TP, una tendenza già riscontrata in tutti i sottocorpora di EPIC (tranne uno).

Nelle prossime sezioni daremo prima spazio alla descrizione dei dati, in modo da delineare le caratteristiche principali degli eventi linguistici selezionati a fare parte del corpus elettronico. In questa descrizione si darà conto soprattutto del numero di eventi linguistici (cioè di testi) presenti nel corpus in rapporto agli attributi più rilevanti, quali la tipologia di partecipanti (interpreti e non interpreti), le sessioni e i tipi di eventi linguistici (così come sono stati classificati secondo la nostra tassonomia), le modalità di emissione dei TP e alcune delle caratteristiche di maggiore interesse dei TP e dei TA: velocità (numero di parole al minuto), durata del tempo di parola e lunghezza (numero di parole effettivamente prodotte).

4.3.1 I partecipanti

4.3.1.1 Gli interpreti

Ai sei interpreti coinvolti nella ricerca sono stati assegnati i seguenti codici identificativi in modo da preservarne l'anonimato (condizione vincolante ed espressa nel consenso informato). Nella seguente Tabella 4.10 sono elencati i codici identificativi e l'indicazione del sesso di ogni interprete (quattro donne e due uomini):

Tabella 4.10 Codici identificativi degli interpreti coinvolti in DIRSI.

Madrelingua italiani (italiano = lingua A; inglese = lingua B)		**Madrelingua inglesi (inglese = lingua A; italiano = lingua B)**	
IT-01	**M**	**UK-01**	**M**
IT-02	**F**		
IT-03	**F**		
IT-04	**F**		
IT-05	**F**		

Nei materiali selezionati per il corpus elettronico sono incluse le performance di tutti gli interpreti ad eccezione di IT-05, ingaggiata per il convegno HIST (presente in archivio, ma non incluso nel corpus). Si tratta a tutti gli effetti di un campione estremamente ridotto, ma in linea con quanto è possibile trovare in altri (rari) esempi descritti in letteratura di studi effettuati con dati raccolti sul

campo. Ovviamente i sei interpreti presentano profili molto diversi tra loro, con un differente grado di esperienza e una non uniforme varietà di ambiti lavorativi prevalenti. Non avendo a disposizione strumenti adeguati per determinare eventuali categorie di differenziazione, possiamo solo rifarci alle caratteristiche salienti del profilo professionale di ciascun soggetto. Una differenziazione rispetto al livello di esperienza potrebbe essere stabilita in base a uno dei criteri di candidatura per essere ammessi come membri della Associazione Internazionale Interpeti di Conferenza (AIIC), ossia l'aver svolto almeno 150 giornate di lavoro secondo gli standard accettati dall'associazione. Così facendo, la popolazione di interpreti coinvolti in DIRSI si ripartirebbe in questo modo: tre interpreti (IT-01, IT-02 e IT-03) sono riconducibili a una categoria che potremmo denominare "*junior interpreter*" avendo svolto alla data delle registrazioni un numero di giornate inferiore alle 150 giornate richieste da AIIC, ma superiore ad almeno la metà di tale quota; i rimanenti interpreti (IT-04, IT-05 e UK-01) rientrano in una categoria che potremmo denominare "*senior interpreter*" con un numero di giornate di lavoro di gran lunga superiore alla quota indicata.

Considerando i convegni selezionati a far parte di DIRSI-C, solamente l'interprete IT-01 è rappresentato in due convegni, mentre tutti gli altri sono rappresentati in un solo convegno. Abbiamo calcolato l'ammontare del tempo di lavoro di ciascun interprete rispetto ai materiali selezionati. Va precisato che questo non equivale alla effettiva suddivisione globale dei turni di lavoro durante i convegni, poiché nel corpus è presente solo una parte dei dati raccolti. Inoltre, è bene ricordare che è proprio sulla base dei turni di lavoro che sono stati definiti i parametri di durata degli eventi linguistici, nonché la loro lunghezza; vi sono casi di TP la cui durata complessiva è stata tale da richiedere un cambiamento di turno da un interprete all'altro. Nella seguente Tabella 4.11 sono riassunti i tempi di lavoro di ciascun interprete per ciascun convegno (le parti incluse in DIRSI-C). Lo stesso dato può essere visto come il tempo di parola complessivo riferito ai TP selezionati a fare parte del corpus:

Tabella 4.11 Durata complessiva dei TA (e dei TP) per interprete e per convegno in DIRSI-C.

Interprete	Convegno	Tempo complessivo di servizio (minuti) in DIRSI-C
IT-01	CFF4 - CFF5	54' – 98' (totale 152')
IT-02	CFF5	129'
IT-03	ELSA	78'
IT-04	ELSA	77'
UK-01	CFF4	125'
Totale		565' (9h 25')

Considerando l'ammontare complessivo del tempo di lavoro degli interpreti per i materiali selezionati, DIRSI-C contiene circa diciannove ore di dati audio (9 ore e 25 minuti di TP a cui si sommano 9 ore e 25 minuti di TA). La durata totale delle registrazioni di ogni convegno è chiaramente superiore, ma sempre più bassa rispetto al tempo di svolgimento globale dell'evento comunicativo vero e proprio (in cui vi sono anche momenti di pausa e interruzioni non inclusi nelle registrazioni).

Rispetto al fattore direzionalità, i dati selezionati per il corpus risultano piuttosto bilanciati, in quanto la somma totale del tempo di lavoro degli interpreti nell'una o nell'altra direzionalità (lingua A o lingua B) è di circa quattro ore e trenta minuti (Tabella 4.12). L'unico interprete con inglese A e italiano B presenta pure una distribuzione temporale equilibrata tra le due lingue di lavoro.

Tabella 4.12 Durata complessiva dei TA per interprete e per convegno in DIRSI-C a seconda della direzionalità.

Convegno	Interprete	Tempo complessivo di servizio (minuti) in DIRSI-C	
		Lingua A	Lingua B
CFF4	IT-01	43'	15'
CFF4	UK-01	58'	67'
ELSA	IT-03	40'	37'
ELSA	IT-04	27'	50'
CFF5	IT-01	38'	60'
CFF5	IT-02	69'	60'
Totale		276' (4h 36')	289' (4h 49')

4.3.1.2 I non-interpreti

Tra tutti i partecipanti non interpreti, ma tipici della situazione comunicativa posta dal convegno sono inclusi nel corpus i seguenti ruoli comunicativi con un diverso numero di eventi linguistici, ovvero di momenti a cui è loro assegnata la facoltà di parola (Tabella 4.13):

Tabella 4.13 Rappresentatività dei partecipanti non interpreti in DIRSI-C per numero totale di eventi linguistici.

Ruolo dei partecipanti	org-it	org-en	Totale
chair	**32**	**0**	**32**
organizer	**12**	**1**	**13**
presenter or lecturer	**10**	**15**	**25**
sponsor	**9**	**0**	**9**

Tabella 4.14 Rappresentatività dei partecipanti non interpreti in DIRSI-C per numero di eventi linguistici nei singoli convegni.

Ruolo dei partecipanti	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
chair	**10**	**11**	**11**	**0**	**0**	**0**
organizer	**4**	**0**	**8**	**0**	**0**	**1**
presenter or lecturer	**2**	**6**	**2**	**5**	**6**	**4**
sponsor	**3**	**1**	**5**	**0**	**0**	**0**

Osservando le due tabelle sopra riportate (Tabella 4.13 e Tabella 4.14) non sorprende l'assenza di ruoli quali *audience* e *discussant*, poiché nel corpus non sono state incluse le sessioni di discussione nella loro interezza (sono infatti considerati solamente gli interventi di apertura, chiusura e le eventuali relazioni esposte in questo tipo di sessione), quali momenti tipici di assegnazione della facoltà di parola anche al pubblico per intervenire con domande, e ai "commentatori" chiamati a fornire una rielaborazione o un contributo aggiuntivo alla relazione principale della sessione in cui sono coinvolti.[10] Un altro dato evidente è l'assenza di interventi in lingua inglese da parte di figure come i moderatori, gli organizzatori (con un solo intervento minimo) e gli sponsor. Una tale distribuzione si spiega se si considera che i convegni in esame sono stati organizzati da soggetti italiani, anche se con il coinvolgimento di colleghi stranieri.

Ulteriori dettagli sui diversi partecipanti ai convegni sono riassunti nella Tabella 4.15 qui di seguito:

[10] Ben diversa sarebbe stata la situazione con il convegno DAYSG (in buona parte trascritto ma non incluso nel corpus): in ogni sessione di discussione, tutti i *discussant* invitati hanno esposto loro stessi delle relazioni, talvolta completamente distaccate dalla relazione principale a cui in teoria erano stati chiamati a riferirsi per offrire un ulteriore approfondimento.

Tabella 4.15 Elenco dei partecipanti non interpreti in DIRSI-C e principali attributi.

Nome	sesso	Numero di interventi	Lingua	Parlante nativo
Mastella	M	12	IT	SÌ
Leonardi,_Barbara	F	11	IT	SÌ
Quattrucci,_Serena	F	5	IT	SÌ
Minicucci	F	4	IT	SÌ
Borgo,_Graziella	F	3	IT	SÌ
Cabrini,_Giulio	M	3	IT	SÌ
Colombo	F	3	IT	SÌ
Giunta	M	3	IT	SÌ
Pieri,_Riccardo	M	3	IT	SÌ
Moran,_Antoinette	F	3	EN	SÌ
Stelmach,_Tiina	F	2	EN	NO
Braggion	M	2	IT	SÌ
Faganelli	M	2	IT	SÌ
Galietta,_Luis	M	2	IT	SÌ
Pandolfini,_Chiara	F	2	IT	SÌ
Ricciardi	M	2	IT	SÌ
Durie,_Peter	M	2	EN	SÌ
Liou,_Theodore	M	2	EN	SÌ
Rosenfeld,_Margaret	F	2	EN	SÌ
Alberti	M	1	IT	SÌ
Cabrini	M	1	IT	SÌ
Fabrizio,_Raffaele	M	1	IT	SÌ
Giordano_Conti	M	1	IT	SÌ
Ibba,_Rossella	F	1	IT	SÌ
NA	M	1	IT	SÌ
Cross,_Anna	F	1	EN	SÌ
Jansone,_Anda	F	1	EN	NO
Mastella	M	1	EN	NO
Sangers,_Sandrina	F	1	EN	NO
Schoemaker,_Freke	F	1	EN	NO

Si noti la presenza di partecipanti che hanno utilizzato l'inglese come L2, mentre i parlanti italofoni sono tutti madrelingua. In totale sono inclusi 16

partecipanti uomini e 14 partecipanti donne. Coloro che presentano un numero elevato di eventi linguistici hanno ricoperto il ruolo di moderatore o di organizzatore. Inoltre, si tenga presente che alcuni dei partecipanti al convegno CFF4 sono rappresentati anche nel convegno CFF5.
La provenienza dei partecipanti è varia per i parlanti anglofoni (sei diversi paesi, di cui solamente due con inglese come lingua ufficiale), mentre tutti i partecipanti italofoni provengono dall'Italia (Tabella 4.16). In realtà, alcuni dei relatori anglofoni hanno iniziato il loro intervento pronunciando qualche parola in italiano (stentato), quale forma di *captatio benevolenziae* nei confronti del pubblico in ascolto.

Tabella 4.16 Rappresentatività dei paesi di provenienza dei partecipanti in DIRSI-C per numero di eventi linguistici.

Paese di provenienza	org-it	org-en	Totale
Estonia	0	2	2
Great Britain	0	1	1
Italy	63	1	64
Latvia	0	1	1
The Netherlands	0	2	2
USA	0	9	9

4.3.2 Macrostruttura e microstruttura del convegno

4.3.2.1 Le sessioni in DIRSI-C

Partendo dalla macrostruttura dei convegni e, in particolare, dalle sessioni incluse nel corpus, DIRSI-C presenta la seguente configurazione (Tabella 4.17 e Tabella 4.18). Le due tabelle contengono dati quantitativi sul numero di eventi linguistici disponibili in DIRSI-C per ogni tipo di sessione:

Tabella 4.17 Rappresentatività delle sessioni in DIRSI-C per numero di eventi linguistici.

Sessione	org-it	org-en	Totale
opening	21	0	21
presentation	27	15	42
discussion	8	0	8
closing	7	1	8

Tabella 4.18 Rappresentatività delle sessioni in DIRSI-C per numero di eventi linguistici nei singoli convegni.

Sessione	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
opening	5	6	10	0	0	0
presentation	9	5	13	5	5	5
discussion	5	0	3	0	0	0
closing	0	7	0	0	1	0

Dalla distribuzione degli eventi linguistici secondo i diversi tipi di sessione, si nota la maggiore rappresentatività della sessione di presentazione, seguita dalla sessione di apertura e dalle altre due sessioni (discussione e conclusioni). Si ricordi che dalle sessioni di discussione sono stati selezionati solo alcuni tipi di eventi linguistici, per questo la quantità di eventi linguistici inclusi nel corpus e provenienti da tale tipo di sessione è così ridotta. Dall'altra parte, la sessione di chiusura apporta un numero esiguo di eventi linguistici perché ad essa è sempre stato dedicato uno spazio notevolmente ridotto rispetto all'intero programma dei tre convegni inclusi nel corpus. Questo è stato rilevato spesso anche per gli altri convegni, in cui l'esiguo tempo a disposizione e la stanchezza dei partecipanti hanno contribuito a snellire i lavori in questa parte conclusiva dell'intero evento comunicativo. Il caso di ELSA è emblematico, in quanto nella restituzione finale di ciò che i partecipanti avevano elaborato all'interno dei gruppi di lavoro (tenuti nel pomeriggio, dopo la sessione plenaria della mattina) le relazioni hanno una durata alquanto ridotta.

4.3.2.2 Gli eventi linguistici in DIRSI-C

Scendendo al livello della microstruttura dei convegni, i dati inclusi nel corpus presentano la seguente distribuzione:

Tabella 4.19 Rappresentatività dei tipi di eventi linguistici in DIRSI-C.

Evento linguistico	org-it	org-en	Totale
floor allocation	14	0	14
opening-closing remarks	32	0	32
paper or lecture	11	14	25
procedure or housekeeping announcements	6	1	7
comment	0	1	1

Tabella 4.20 Rappresentatività dei tipi di eventi linguistici nei singoli convegni di DIRSI-C.

Evento linguistico	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
floor allocation	3	6	5	0	0	0
opening-closing remarks	11	4	17	0	0	0
paper or lecture	3	5	3	4	6	4
procedure or housekeeping announcements	2	3	1	0	0	1
comment	0	0	0	1	0	0

Come si vede dalle due tabelle sopra riportate (Tabella 4.19 e Tabella 4.20) vi è un alto numero di interventi di apertura-chiusura (32 occorrenze, tutte dal sottocorpus di TP in italiano), seguito dalle relazioni (*paper or lecture*, con 25 occorrenze) e dagli interventi di assegnazione della facoltà di parola (14 occorrenze). Sono incluse anche 7 occorrenze di eventi linguistici classificati come "procedurali" e una sola occorrenza di "commento". Si consideri che la classificazione dei diversi tipi di eventi linguistici è stata eseguita sulla base della funzione dominante riscontrata in ciascun "testo". Questo significa che vi possono essere interventi "misti" nei quali viene aperta o chiusa una sessione e, al contempo, viene assegnata la facoltà di parola. Spesso la chiusura di una sessione e l'apertura della sessione successiva sono riconducibili allo stesso evento linguistico, il che spiega la necessità di accorpare le due tipologie nella tassonomia da noi utilizzata.

Nelle prossime sezioni saranno presentate nel dettaglio le caratteristiche degli eventi linguistici inclusi in DIRSI-C.

4.3.3 Caratteristiche degli eventi linguistici in DIRSI-C

Le principali caratteristiche degli eventi linguistici inclusi nel corpus sono state esaminate estraendo le informazioni rilevanti (cioè gli attributi) inserite nell'*header* di ciascuna trascrizione (§4.2.3.3, Tabella 4.5). Il recupero di queste informazioni è stato effettuato in maniera semiautomatica, sfruttando quindi la natura *machine-readable* dei dati in DIRSI-C.

4.3.3.1 Modalità di emissione del TP

All'interno della categoria "modalità di emissione" abbiamo fatto rientrare il grado di oralità dei diversi eventi linguistici, distinguendo tra spontaneo (*impromptu*), semispontaneo o preparato (*mixed*) e letto (*read*), nonché l'eventuale uso di supporti audiovisivi.

4.3.3.1.1 Grado di oralità

Tabella 4.21 Rappresentatività dei diversi gradi di oralità in DIRSI-C per numero di eventi linguistici.

Modalità di emissione TP	org-it	org-en	Totale
impromptu	56	3	59
mixed	7	13	20
read	0	0	0

Tabella 4.22 Rappresentatività dei diversi gradi di oralità in DIRSI-C per numero di eventi linguistici nei singoli convegni.

Modalità di emissione TP	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
impromptu	19	15	22	1	1	1
mixed	0	3	4	4	5	4
read	0	0	0	0	0	0

Osservando le due tabelle riportate in questa sezione (Tabella 4.21 e Tabella 4.22) si nota l'assenza di interventi esposti attraverso la lettura integrale di un testo. Ritroviamo 20 eventi linguistici nella categoria "mixed" corrispondenti alle relazioni (paper or lecture), alcune delle quali sono state emesse senza una preparazione previa e in modalità completamente spontanea. Molti sono gli interventi totalmente spontanei, tra cui gli interventi per l'assegnazione della facoltà di parola, e gli interventi di apertura/chiusura.

4.3.3.1.2 Uso di supporti audiovisivi

Per quanto riguarda l'uso di supporti audiovisivi, è noto come l'uso delle presentazioni power point sia diventato quasi una regola obbligatoria per chi si

accinge a tenere una relazione a un convegno. I lucidi sono raramente usati, mentre talvolta i relatori fanno riferimento a documentazione scritta a disposizione del pubblico. Nei dati inclusi in DIRSI-C l'eventuale uso di supporti audiovisivi riguarda sempre le presentazioni in power point:

Tabella 4.23 Rappresentatività dell'uso di supporti audiovisivi in DIRSI-C per numero totale di eventi linguistici.

Uso di supporti audiovisivi	org-it	org-en	Totale
no	45	2	57
yes	9	13	22

Tabella 4.24 Rappresentatività dell'uso di supporti audiovisivi in DIRSI-C per numero di eventi linguistici nei singoli convegni.

Uso di supporti audiovisivi	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
no	8	15	22	0	1	1
yes	2	3	4	4	5	4

Il dato di vero interesse per questo attributo non è tanto l'assenza di supporti audiovisivi per un alto numero di eventi linguistici. Questo è comprensibile, poiché come è stato già sottolineato sono le relazioni gli interventi in cui è plausibile (e accettabile all'interno delle dinamiche del convegno) l'eventuale uso di tali supporti. Le 22 occorrenze registrate nel corpus confermano questa prassi per le relazioni, in linea con quanto è avvenuto anche negli altri convegni immagazzinati in DIRSI-MA.

4.3.3.2 Velocità di eloquio

La velocità di emissione dei TP è stata calcolata mettendo in rapporto il numero di parole e il tempo a disposizione in ciascun evento linguistico. I valori soglia fissati per differenziare le tre categorie di velocità "alta", "media" e "bassa" sono stati presentati precedentemente (§4.2.3.3, Tabella 4.5). Osservando i dati inclusi nel corpus si ottiene la seguente distribuzione tra le varie categorie di velocità per le due diverse lingue (Tabella 4.25 e Tabella 4.26):

Tabella 4.25 Rappresentatività della velocità di eloquio in DIRSI-C per numero totale di eventi linguistici.

Speed (parole al minuto)	org-it	org-en	Totale
high (>120)	20	11	31
medium (100-120)	23	3	26
low (< 100)	20	2	22

Tabella 4.26 Rappresentatività della velocità di eloquio in DIRSI-C per numero di eventi linguistici nei singoli convegni.

Speed (parole al minuto)	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
high (>120)	5	8	7	5	2	4
medium (100-120)	10	6	7	0	3	0
low (< 100)	4	4	12	0	1	1

Confrontando i risultati per le due lingue, si nota la prevalenza di "testi" inglesi emessi a una velocità superiore alle 120 parole al minuto, mentre i "testi" italiani appaiono più distribuiti fra le tre categorie. Ad ogni modo, il numero totale di eventi linguistici emessi a una velocità inferiore alle 100 parole al minuto è decisamente più basso rispetto al numero di eventi linguistici prodotti a velocità superiori.
Abbiamo calcolato la velocità media di tutti gli eventi linguistici presenti in ciascun sottocorpus DIRSI-C, sia nei sottocorpora di TP, sia nei sottocorpora di TA (Tabella 4.27):

Tabella 4.27 Velocità media (numero di parole al minuto) in DIRSI-C.

sottocorpus	numero di testi	Velocità media (parole al minuto)
ORG-IT	63	110
INT-IT-EN	61*	120
ORG-EN	16	129
INT-EN-IT	16	117

*Nel calcolo della media non sono stati considerati i due casi in cui l'interprete non ha tradotto il TP.

Dai dati presentati nella Tabella 4.27 emerge un risultato per certi versi inatteso: la velocità media dei TP italiani (110 p/m) è di gran lunga inferiore alla velocità media dei TP inglesi (129 p/m). Anche se si considerano solo i TP italiani emessi ad una velocità media e alta, escludendo quindi i 20 "testi" emessi a una velocità bassa, la media si attesta a 121 parole al minuto, rimanendo inferiore di otto punti rispetto al dato dei TP inglesi (di cui non ci sono occorrenze di eventi linguistici emessi a una velocità bassa e solamente due a velocità media). Una tendenza simile è stata riscontrata anche in EPIC, anche se i livelli di velocità sono generalmente molto più elevati: la velocità media dei TP inglesi è di circa 156 parole al minuto, mentre la velocità media dei TP italiani è di 130 parole al minuto (i dati in EPIC comprendono anche la velocità media dei TP spagnoli che arrivano a 152 parole al minuto).

Dall'altra parte, la velocità media dei TA nelle due lingue considerate in DIRSI-C è piuttosto simile: 120 parole al minuto nei TA in inglese e 117 parole al minuto nei TA in italiano. La stessa tendenza è confermata anche nei dati EPIC, dove la cabina inglese lavora a una media di 132 parole al minuto e la cabina italiana lavora a una media di 124 parole al minuto. I valori ottenuti in EPIC sono generalmente più elevati, ma rispecchiano un andamento simile rispetto ai valori ottenuti in DIRSI-C.

4.3.3.3 Durata (tempo di parola)

I valori soglia impiegati per suddividere gli eventi linguistici in tre categorie di durata (Tabella 4.6) sono stati illustrati nella prima parte di questo capitolo (§4.2.3.3). Si ricordi che il punto di osservazione adottato ai fini della classificazione dei dati è quello dell'interprete (oltre che un punto di osservazione generale ed esterno all'evento comunicativo). Questo significa che i TP "suddivisi" tra i due interpreti in servizio sono stati "spezzati" e trattati come "testi" (eventi linguistici) autonomi. Non a caso il valore soglia oltre il quale è da applicare la categoria "long duration" è 1800 secondi (cioè 30 minuti). In questo senso, laddove un interprete abbia tradotto continuativamente un TP per oltre 30 minuti si può parlare di un TP di lunga durata. Diversamente, a seconda di come il TP nella sua interezza sia stato gestito dagli interpreti in servizio si potranno avere TP di durata lunga, media o breve in riferimento allo stesso TP globale. Per fare un esempio, nel convegno CFF4 la prima relazione ha avuto una durata complessiva di poco meno di 50 minuti, e sarebbe pertanto da classificare come evento linguistico di lunga durata. Tuttavia, la prima parte dell'intervento è stata tradotta da IT-01 (26 minuti), mentre la seconda parte è stata tradotta dall'altro interprete UK-01 (23 minuti). Di conseguenza, nella strutturazione dei materiali all'interno del corpus le due parti dello stesso TP risultano suddivise (poiché da riferire a due diversi interpreti) e sono classificate

come TP di durata media. Dalla denominazione del file è comunque possibile risalire al fatto che si tratta di uno stesso TP suddiviso in due diverse parti (§4.1).

Tabella 4.28 Rappresentatività della durata dei TP in DIRSI-C per numero totale di eventi linguistici.

Duration	org-it	org-en	Totale
long (> 30')	1	3	4
medium (15'-30')	4	7	11
short (< 15')	58	6	64

Tabella 4.29 Rappresentatività della durata dei TP in DIRSI-C per numero di eventi linguistici nei singoli convegni.

Duration	org-it			org-en		
	CFF4	ELSA	CFF5	CFF4	ELSA	CFF5
long (> 30')	0	0	1	1	0	2
medium (15'-30')	1	2	1	3	2	2
short (< 15')	18	16	24	1	4	1

I risultati ottenuti analizzando i materiali inclusi nel corpus rispecchiano un profilo riscontrato anche in altri convegni, in quanto vi è un nucleo di interventi di durata media (tra i 15 e i 30 minuti) e lunga (oltre i 30 minuti) attorno al quale si distribuiscono tanti interventi di durata decisamente inferiore (al di sotto dei 15 minuti). Pur avendo selezionato solamente alcuni tipi di eventi linguistici e sessioni nella creazione di DIRSI-C, questa caratteristica si è mantenuta costante.

4.3.3.4 Lunghezza (numero di parole)

Assieme agli altri attributi che sono stati presentati, anche i valori soglia fissati per stabilire tre diverse categorie di lunghezza dei testi (Tabella 4.7) sono stati illustrati precedentemente (§4.2.3.3). La stessa osservazione sulla suddivisione dei TP da parte degli interpreti esposta nella sezione precedente si deve considerare nel calcolo della lunghezza dei TP. Vale la pena sottolineare nuovamente che nel nostro caso il numero di parole non corrisponde al numero di *token* che si ottiene utilizzando strumenti di linguistica computazionale (§1.3.4.1). Questo è dovuto al fatto che il numero di parole è stato calcolato

utilizzando la funzione apposita presente in TextPad, il programma di videoscrittura usato per gestire le trascrizioni (lo stesso si sarebbe ottenuto usando la stessa funzione disponibile nel programma MS Word). La particolarità di tale calcolo è che le parole apostrofate sono considerate un tutt'uno con la parola a cui sono unite mediante l'apostrofo (come nei casi presenti in questa stessa ultima frase, "tutt'uno" e "l'apostrofo"), e quindi conteggiate come una sola parola. Per contro, la tokenizzazione "spezza" questa unione e conteggia due *token* laddove era stata calcolata la presenza di una sola parola. Si tratta di una differenza importante che non può essere sottovalutata in fase di analisi.

Tabella 4.30 Rappresentatività della lunghezza dei TP in DIRSI-C per numero totale di eventi linguistici.

Speech length (numero di parole)	**org-it**	**org-en**	**Totale**
long (> 3300)	**3**	**3**	**6**
medium (1650-3300)	**2**	**7**	**9**
short (< 1650)	**58**	**6**	**64**

Tabella 4.31 Rappresentatività della lunghezza dei TP in DIRSI-C per numero di eventi linguistici nei singoli

Speech length (numero di parole)	**org-it**			**org-en**		
	CFF4	**ELSA**	**CFF5**	**CFF4**	**ELSA**	**CFF5**
long (> 3300)	**1**	**1**	**1**	**1**	**0**	**2**
medium (1650-3300)		**1**	**1**	**3**	**2**	**2**
short (< 1650)	**18**	**16**	**24**	**1**	**4**	**1**

Osservando la distribuzione dei dati nelle due tabelle sopra riportate (Tabella 4.30 e Tabella 4.31) notiamo che essa riflette fedelmente quanto illustrato in merito alla durata in termini di tempo degli stessi eventi linguistici. Se così non fosse, sarebbe stato necessario tarare diversamente i valori soglia che sono stati stabiliti al fine di classificare i materiali raccolti nell'archivio multimediale e inseriti nel corpus.

Capitolo 5

I corpora di interpretazione tra ricerca e didattica

5.1 Potenzialità di ricerca

I dati disponibili nei due corpora EPIC e DIRSI-C, illustrati nei precedenti capitoli, sono già stati oggetto di diversi studi e si prestano ad innumerevoli esplorazioni future. Uno dei vantaggi maggiori è dato dalla possibilità di estrarre dai corpora tutte le occorrenze di particolari fenomeni in maniera automatica, ricavandoli dal tipo e dal livello di annotazione effettuati durante la creazione di ciascuna risorsa.[1] Vale la pena precisare che i risultati ottenuti da eventuali liste di frequenza necessitano di ulteriori analisi e riflessioni da parte del ricercatore, in un processo di osservazione, discussione e interpretazione che per sua natura si nutre del tempo e dell'intelletto di chi si dedica alla ricerca.

Riprendendo i tre livelli delle convenzioni seguite per le trascrizioni, si possono considerare ricerche di tipo linguistico, paralinguistico e, in base ai metadati inclusi nell'*header* impostato per ogni corpus, di tipo extralinguistico. Grazie poi all'annotazione grammaticale e alla lemmatizzazione, tali operazioni di interrogazione del corpus diventano più agevoli e aprono a un ventaglio di opzioni più ampio.

Uno dei primi studi condotti su EPIC riguarda la densità lessicale (Sandrelli & Bendazzoli 2005, Russo et al. 2006, Sandrelli et al. 2010). Ispirato da un'analisi simile svolta da Laviosa (1998b) su un corpus comparabile di testi narrativi originali e tradotti, in lingua inglese, l'obiettivo dello studio era

[1] Ulteriori annotazioni sono possibili, in quanto l'impostazione di tutto il materiale raccolto è tale da consentire una buona flessibilità nella rielaborazione del materiale stesso, ampliando in questo modo le potenzialità d'analisi e i percorsi di ricerca. Liste di frequenza, densità lessicale, concordanze e collocazioni sono tra gli esempi classici di impiego della linguistica dei corpora, analisi che si possono finalmente svolgere anche su trascrizioni di discorsi orali e interpretati in simultanea.

verificare se i risultati ottenuti da Laviosa potessero essere riscontrati anche nei TP e TA presenti in EPIC, cioè discorsi orali originali e interpretati. Questo tipo di analisi è stato possibile grazie all'estrazione automatica delle liste di parole lessicali e di parole grammaticali contenute in ciascun sottocorpus di EPIC, con le quali sono state calcolate la densità lessicale in ciascun sottocorpus e il grado di varietà lessicale nelle cosiddette *list head*, cioè le porzioni di corpus costituite dalle 100 parole più frequenti.

Un'altra analisi resa possibile dal particolare tipo di annotazione applicato alle parole lasciate incomplete e alle parole mal pronunciate in EPIC ha permesso di delineare un profilo di queste due disfluenze in interpretazione simultanea (Bendazzoli et al. 2011). Anche in questo caso, le occorrenze di tutte le parole caratterizzate dalle disfluenze prese in esame sono state estratte automaticamente da ciascun sottocorpus di EPIC. A questa fase quantitativa è poi seguito uno studio qualitativo per ogni occorrenza, in modo da ricavarne le caratteristiche più rilevanti. Gli stessi risultati sono stati anche messi in rapporto ad alcuni parametri inclusi nell'*header*, quali l'argomento trattato, la modalità di presentazione del TP e la velocità di eloquio (Sandrelli et al. 2007), applicando test statistici che hanno messo in luce la significatività o meno delle differenze riscontrate.

Infine, in un contributo di tipo più metodologico (Sandrelli & Bendazzoli 2006) è stata valutata la correttezza dell'annotazione grammaticale e della lemmatizzazione effettuata dai *tagger* impiegati (Treetagger e Freeling). Tale confronto è stato esteso a Grampal, un secondo *tagger* per il materiale spagnolo (Moreno Sandoval & Goñi 1995, Guirao & Moreno Sandoval 2004) con il quale è stata effettuata una annotazione a parte.[2]

Per quanto riguarda DIRSI-C, nell'ambito del programma di dottorato in cui è stato creato sono stati condotti tre studi preliminari, che saranno ripetuti su più ampia scala anche in EPIC. Il primo concerne una misurazione classica nella linguistica computazionale, ovvero la *type-token ratio*, per valutare la varietà lessicale. Il calcolo è stato effettuato per ciascuno dei quattro sottocorpora di cui si compone DIRSI-C grazie al formato elettronico in cui sono disponibili le trascrizioni annotate e indicizzate. Nel secondo studio è stata presa in esame la produzione verbale nei TP rispetto alla produzione verbale nei TA. In altre parole, è stato verificato se la tendenza generale alla riduzione del numero di parole prodotte dagli interpreti rispetto alla quantità di parole prodotte dagli oratori originali si verificasse sempre e indipendentemente dalla durata del TP. I dati inclusi nell'*header* di ciascuna trascrizione hanno consentito di affinare l'osservazione a livello dei singoli eventi linguistici, oltre a considerare il

[2] Questa annotazione è stata effettuata da José María Guirao (Università di Granada) e Antonio Moreno Sandoval (Laboratorio di Linguistica Informatica, Università Atuonoma di Madrid) con i quali è in corso una fruttuosa collaborazione.

risultato globale per ciascun sottocorpus. Infine, nel terzo esempio di analisi è stato preso in esame il segnale discorsivo *so* nei TA in inglese, al fine di comprovare se questo fosse stato prodotto dagli interpreti in risposta a un'unità di significato equivalente nel TP o se fosse il risultato di una aggiunta. In questo caso, dall'analisi quantitativa con l'estrazione automatica di tutte le occorrenze del segnale discorsivo considerato, si è poi passati a svolgere un'analisi qualitativa con l'ausilio della interfaccia in cui i materiali sono disponibili allineati sia a livello testo-suono sia a livello TP-TA.

Oltre alle esplorazioni già compiute e qui solo brevemente menzionate, per entrambi i corpora si aprono innumerevoli percorsi di indagine, in prospettiva sia parallela sia comparabile. Ad esempio, si potranno esaminare le caratteristiche salienti della "lingua di conferenza" (Riccardi 1997, Garzone & Viezzi 2001) e confrontarle con i tratti tipici della "lingua interpretata" attraverso l'estrazione di collocazioni, bigrammi e trigrammi, così come il calcolo di indici relativi all'articolazione paratattica e ipotattica, alla varietà lessicale, nonché alla frequenza e varietà dei segnali discorsivi. Similmente, facendo eco al concetto di *translational norms* nell'interpretazione simultanea (Schjoldager 1995), l'osservazione potrà essere mirata a fenomeni che sono manifestazioni ricorrenti di particolari strategie traduttive e processi cognitivi attivati dagli interpreti, quali le aggiunte, l'anticipazione, le omissioni, la traduzione dei nomi propri e dei numeri; o ancora ci si potrà concentrare su aspetti di natura sociolinguistica e pragmatica, come la gestione della *politeness* e delle strategie di *hedging*, dei pronomi personali, l'uso dei segnali discorsivi e della deissi.

L'elenco potrebbe continuare all'infinito,[3] così come continuo è tuttora l'impegno richiesto per espandere ulteriormente i corpora a disposizione, arricchendoli di altri materiali e annotazioni. Gli archivi multimediali associati ad EPIC e DIRSI-C rappresentano fonti preziose di dati che possono essere utilizzati in altri studi, la cui realizzazione contribuisce al contempo alla crescita dei due corpora, in quanto nuove trascrizioni potranno essere aggiunte in futuro. Questa esperienza è già stata compiuta con successo, coinvolgendo i laureandi della Scuola Superiore di Lingue Moderne per Interpreti e Traduttori di Forlì che hanno avuto la possibilità di usufruire dei materiali EPIC in archivio per le loro tesi di laurea (Russo 2010).

[3] Sono da includere anche le ricerche di tipo sperimentale, per le quali i TP del Parlamento europeo sono già stati utilizzati con profitto (Viezzi 1999, Donato 2003).

5.2 Potenzialità didattiche

Oltre che per le attività di ricerca, gli archivi multimediali e i corpora possono essere sfruttati nella formazione dei futuri interpreti e nell'apprendimento delle lingue straniere. La disponibilità di materiali autentici e strutturati secondo uno schema di classificazione coerente ha senza dubbio un valore notevole per chiunque si voglia avvicinare al mondo dell'interpretazione (Messina 2001, Sandrelli & de Manuel 2007). Tra gli esempi di archivi e *speech repositories* che sono stati menzionati (§1.2.1), merita particolare attenzione la banca dati *Marius* (de Manuel 2003b), sviluppata presso l'Università di Granada a partire anche da materiali tratti dalle istituzioni europee.[4]

Uno dei vantaggi principali, sia per i formatori sia per i discenti, sta nel poter reperire comodamente diversi esempi di eventi linguistici, raggruppati in base a parametri come la velocità di eloquio, la durata, l'argomento, la modalità di esposizione, e tutte le altre caratteristiche incluse nell'*header* di ogni trascrizione. Inoltre, vi è la possibilità di "confrontare" la propria prestazione con quella dei professionisti ingaggiati nella situazione reale da cui i materiali provengono. Tale confronto non deve essere concepito come l'ascolto della versione "corretta", ma risulta utile in quanto stimola l'individuazione di strategie specifiche che sono state adottate dagli interpreti professionisti.

L'impiego dei corpora a fini didattici potrebbe dimostrarsi valido non solo in classe, ma anche nelle attività di autoapprendimento. A questo proposito, i materiali testuali e multimediali di EPIC e DIRSI-C, con i loro formati "flessibili", sono di certo compatibili con programmi sviluppati appositamente per promuovere l'autoapprendimento. Ne è un esempio Black Box (Sandrelli 2002, 2003a, 2003b), una *suite* di programmi in cui si possono impostare veri e propri pacchetti didattici con dati multimediali e testuali, da arricchire con ulteriori collegamenti ipertestuali.

La stessa attività di trascrizione in cui i laureandi della SSLMIT sono stati coinvolti nel realizzare le loro tesi di laurea ha aumentato la loro consapevolezza sullo stile di eloquio tipico di un contesto istituzionale come il Parlamento europeo.[5] Più in generale, come già suggerito da Aston (1997) per la glottodidattica, pur occupandosi di un campione di dati limitato in termini quantitativi per ovvie ragioni, l'applicazione della metodologia CIS da parte degli interpreti in formazione non può che andare a loro vantaggio, in quanto li

[4] Questo è un ulteriore segnale dei numerosi vantaggi che si possono trarre dall'uso di dati provenienti dai contesti comunitari, primo fra tutti il Parlamento europeo (Bendazzoli 2010).

[5] Alcuni materiali sono anche stati impiegati nell'ambito di un corso *Sound Perception Trainer* (Kaunzner 1997, Kaunzner & Gianni 1997) con l'obiettivo di migliorare la pronuncia e di far sì che gli studenti familiarizzassero con il "gergo" europarlamentare.

espone ai linguaggi specialistici e alle prassi comunicative tipiche degli eventi mediati da interpreti.

Un'altra esperienza interessante sullo sviluppo di attività didattiche ricavate da EPIC riguarda la gestione del passato prossimo nella simultanea dall'italiano verso l'inglese e lo spagnolo (Sandrelli 2010), lingue che prevedono una gestione dei tempi verbali diversa rispetto all'italiano in presenza di determinati indicatori temporali. Grazie all'estrazione automatica di tutte le occorrenze dei verbi espressi al passato prossimo, è stato possibile innanzitutto studiarne l'effettiva resa nelle due lingue di arrivo (Graupera 2009). Non deve sorprendere se questa si è rivelata prevalentemente corretta, in quanto per le lingue considerate in EPIC si riscontra sempre una direzionalità B>A, cioè dalla lingua straniera alla lingua materna degli interpreti. Proprio sulla base di questo risultato «we maintain that corpora of speeches interpreted by native speakers (i.e. by professionals working into their A language) could be particularly useful to trainees learning to interpret into their B language» (*ibid*., p. 86). Sulla base di questo principio, altri tratti linguistici potranno essere esaminati in dettaglio, come ad esempio la resa del congiuntivo dall'italiano in inglese (Mead 2002b) e viceversa (Viezzi 2002).

Dall'unione dei percorsi pedagogici menzionati, uno degli sviluppi futuri dei progetti EPIC e DIRSI prevede la creazione di unità didattiche per l'interpretazione, mirate per esempio al consolidamento della comprensione e produzione delle formule procedurali, di apertura e di chiusura solitamente impiegate nei convegni.[6] Con tutti i dati a disposizione, sarà possibile offrire molteplici esempi dal vivo, oltre che batterie di esercizi focalizzati su abilità specifiche.

Infine, le potenzialità didattiche dei CIS potrebbero essere estese più in generale all'apprendimento linguistico, sulla scia dei progressi compiuti in svariati ambiti della *Corpus Linguistics* (Facchinetti 2007, Hidalgo et al. 2007). Questo è vero soprattutto per chi è interessato a rafforzare le proprie competenze comunicative di ascolto e produzione orale all'interno di una particolare comunità linguistica, come coloro che sono chiamati a tenere presentazioni in occasioni di convegni.[7] Così come la traduzione è stata inclusa da tempo tra le attività di formazione in ambito linguistico (Malmkjaer 1998, Cook 2010), senza puntare quindi specificatamente alla formazione di traduttori veri e propri, anche le attività didattiche tradizionalmente pensate per gli interpreti in formazione potrebbero trovare spazio in ambiti pedagogici di più ampio respiro. A questo proposito, la validità dell'approccio che potremmo inquadrare come *interpretation in language teaching* è già stata dimostrata in alcuni casi

[6] A tal riguardo, ci si potrà basare anche su ricerche condotte in contesti comunicativi non mediati da interpreti (cfr. Rowley-Jolivet & Carter-Thomas 2005).

[7] Si vedano le attività didattiche proposte da Bayne (2005) riconducibili alla sessione poster di un convegno.

(Zannirato 2008) e la disponibilità di nuove risorse linguistiche provenienti dai CIS non potrà che moltiplicare le intersezioni tra settori didattici affini che potranno così beneficiare l'uno dell'altro.

Considerazioni finali

L'obiettivo generale che ci eravamo posti all'inizio di questo lavoro era giungere alla definizione di una metodologia specifica per applicare pienamente il *corpus-based approach* allo studio dell'interpretazione simultanea, abbracciando comunque tutte le altre modalità di interpretazione riconducibili ai *Corpus-based Interpreting Studies* (CIS). Lo sviluppo di questa branca della Traduttologia è stato per certi versi più lento rispetto alla sua controparte "scritta", ovvero i *Corpus-based Translation Studies* (CTS) riguardanti i testi di traduzione. I motivi di tale divario sono tanto di ordine metodologico quanto di ordine pratico, e vanno dalla difficoltà di registrare eventi comunicativi mediati da interpreti (con il loro consenso) alle questioni sulla rappresentazione in forma scritta del parlato, cioè la trascrizione; dalla conoscenza delle tecnologie disponibili per interfacciare tra loro dati di diversa natura (testuali e multimediali) alla disponibilità di strumenti di analisi computazionali in grado di gestire le tante peculiarità dei testi di partenza e di arrivo, altrimenti detti discorsi originali e interpretazioni.

Nel primo capitolo è stata discussa l'applicazione del *corpus-based approach* agli studi sull'interpretazione, fornendo innanzitutto una definizione di *corpus* che ne ha messo in luce le caratteristiche fondamentali in riferimento all'ambito disciplinare della linguistica computazionale. Abbiamo considerato, pertanto, non solo l'aspetto quantitativo e rappresentativo (sulla base di criteri di inclusione e di esclusione) dei materiali costitutivi di un corpus, ma anche il formato in cui sono resi disponibili gli stessi materiali. Tale formato dovrebbe essere *machine-readable*, ovvero tale da poter utilizzare il più alto numero di applicazioni informatiche in grado di elaborare i dati in maniera semi-automatica dal computer.

Successivamente, sono state descritte diverse tipologie di corpus, a seconda delle caratteristiche dei dati considerati e della rappresentatività che si vuole ottenere, riservando spazio in particolare agli *spoken corpora*. Già in questa fase è stato possibile evidenziare le maggiori difficoltà pratiche e metodologiche insite nella creazione di un corpus orale, difficoltà che possono diventare veri e propri ostacoli quando ci si propone di occuparsi di oralità mediata, cioè di interpretazione. Nella seconda parte di questo stesso capitolo

sono state descritte dettagliatamente le principali sfide metodologiche tipiche dei CIS e presenti in ciascuna delle tappe implicate nella creazione di un *machine-readable interpreting corpus*: corpus *design*, con particolare riferimento alla struttura e alla rappresentatività del corpus; raccolta dei dati, considerando i diversi tipi e gradi di accesso ai dati sul campo, l'uso del consenso informato per la registrazione e i requisiti tecnici per realizzarla nella pratica; trascrizione, ovvero la trasformazione dei dati da entità evanescenti a entità stabili e misurabili (procedimento che comporta non poche questioni teoriche e pratiche); codifica e annotazione, ossia l'arricchimento delle trascrizioni con informazioni sulla lingua rappresentata nel corpus e il modo in cui strutturarla perché sia analizzabile in maniera assistita; allineamento (testo-suono e TP-TA); accessibilità e distribuzione del corpus.

Nel secondo capitolo sono state presentate alcune delle iniziative più rilevanti all'interno dei CIS, che ne hanno segnato l'evoluzione, differenziando però i prodotti di tali ricerche in tre grandi gruppi: i corpora manuali, cioè raccolte di dati rappresentative, ma prive di annotazioni e codifica e che pertanto possono essere analizzate ancora solo secondo metodi tradizionali e senza l'ausilio di specifici programmi informatici; i corpora elettronici non accessibili (fino a questo momento) al resto della comunità scientifica; i corpora elettronici e liberamente accessibili, tra cui i risultati ottenuti dai due progetti di ricerca che sono stati approfonditi nei due capitoli successivi: il corpus EPIC (cap. 3) e il corpus DIRSI-C (cap. 4).

Il corpus EPIC con il suo archivio multimediale è di fatto uno dei primi corpora elettronici ad essere stato messo a disposizione nel campo degli *Interpreting Studies*. Si tratta di un corpus trilingue di TP in italiano, inglese e spagnolo con i rispettivi TA in tutte le direzioni possibili tra le lingue menzionate. Il contesto di provenienza dei materiali è dato dalle sedute plenarie del Parlamento europeo, un contesto unico nel suo genere per la mole di attività traduttiva e per la configurazione linguistica che ne deriva (Sunnari 1997, 1999; Cosmai 2003; Bendazzoli 2010). La possibilità di accedere ai dati da un canale esterno (le trasmissioni satellitari della rete EbS, e ora anche da Internet), nonché il permesso d'uso concesso dai responsabili dei servizi audiovisivi del PE per attività di ricerca e didattica (non commerciali) hanno orientato la scelta dei dati su cui concentrare gli sforzi per questa prima iniziativa CIS. Il particolare contesto di provenienza ha determinato un profilo specifico dei TP e delle condizioni di lavoro degli interpreti. Questo significa che per quanto EPIC abbia grandi potenzialità di fornire risultati di ricerca rappresentativi, essi dovranno sempre essere rapportati al contesto specifico di provenienza dei dati stessi.

EPIC comprende registrazioni video dei TP e audio dei TA. Le trascrizioni sono state annotate a livello grammaticale, lemmatizzate e indicizzate secondo lo standard della *Corpus Work Bench* (CWB-CQP). Ad ogni modo, gli stessi

dati sono disponibili in formati estremamente flessibili e compatibili con altre applicazioni, con cui si potrebbero elaborare conformemente ad altri standard. Il corpus è inoltre esplorabile attraverso una interfaccia di ricerca *online*, sviluppata dal personale tecnico della Scuola Superiore di Lingue Moderne per Interpreti e Traduttori dell'Università di Bologna (sede di Forlì). Da tale interfaccia sono per ora accessibili solo le trascrizioni, mentre i dati multimediali sono accessibili dalla rete del Dipartimento SITLeC.

Il corpus elettronico DIRSI-C e l'archivio multimediale da cui sono stati selezionati i materiali che fanno parte del corpus potrebbero essere considerati come una prosecuzione del percorso già compiuto nel creare il precedente corpus, rispetto al quale sono stati apportati ulteriori sviluppi. L'acronimo DIRSI (*Directionality in Simultaneous Interpreting*) mette in risalto il parametro principe che caratterizza il funzionamento di questo corpus, ossia la direzionalità. Nel considerare questo parametro si vuole tenere conto dell'effetto che la lingua di lavoro degli interpreti può avere sulla loro prestazione, a seconda che si tratti della loro lingua A (la lingua materna), di una lingua B (una lingua straniera verso la quale sono in grado di tradurre) o di una lingua C (una lingua straniera da cui sono in grado di tradurre, ma verso cui non fornirebbero una prestazione professionalmente accettabile). È proprio il focus su questo parametro a differenziare DIRSI-C da EPIC. L'elemento di maggior novità è direttamente collegato alla fonte stessa da cui sono stati raccolti i dati per DIRSI, ossia il contesto dei convegni internazionali svolti nel mercato italiano e mediati da interpreti professionisti. Questo ha garantito la disponibilità di prestazioni professionali di interpretazione simultanea tra l'italiano e l'inglese i cui esecutori (gli interpreti) hanno lavorato dalla lingua straniera (lingua B) verso la loro lingua madre (lingua A) e viceversa (cioè da A a B), a differenza di quanto avviene di norma nel contesto delle sedute plenarie del Parlamento europeo, dove gli interpreti lavorano quasi esclusivamente dalle loro lingue B verso la loro lingua A (con le dovute eccezioni, soprattutto in seguito alle diverse fasi di allargamento dell'Unione).

Avendo dovuto operare maggiormente sul campo rispetto alla precedente esperienza, nella prima parte del quarto capitolo abbiamo esposto il nostro "diario di bordo" sulla creazione dell'archivio multimediale e del corpus elettronico DIRSI, entrambi frutto dell'applicazione concreta di tutti gli strumenti metodologici individuati e del confronto con le tante sfide messe a fuoco con EPIC. Nell'archivio sono immagazzinate le registrazioni e i dati raccolti sul campo da 14 convegni internazionali tenuti in Italia tra il 2006 e il 2010 (Tabella 4.1, §4.1), coinvolgendo in totale sei interpreti professionisti (di cui cinque madrelingua italiani e un madrelingua inglese). La procedura di archiviazione dei dati e la successiva creazione del corpus hanno comportato la messa a punto di un sistema di classificazione specifico. A tal fine, è stato

necessario attingere a diversi contributi metodologici, inerenti allo studio e all'analisi delle situazioni comunicative, provenienti da altre discipline, quali l'Etnografia della comunicazione, l'Antropologia del linguaggio, la Sociolinguistica, l'Analisi conversazionale, l'Analisi del discorso e gli stessi Studi sull'interpretazione. Grazie a questa disamina è stato possibile definire gli strumenti operativi più adeguati alla nostra analisi e le scelte metodologiche più consone alla classificazione dei materiali da includere nel corpus. In particolare, l'attenzione è stata posta sui tipi di sessione in cui si strutturano i convegni, i diversi tipi di interventi ratificati (eventi linguistici) prodotti dai partecipanti, nonché i ruoli comunicativi attribuibili a questi ultimi nella situazione comunicativa in questione.

Da una selezione di dati appartenenti a tre dei 14 convegni immagazzinati in archivio è stato creato il corpus elettronico DIRSI-C (§4.3), contenente oltre 135.000 parole in totale, suddivise in quattro sottocorpora: uno con i TP italiani, uno con i TP inglesi, uno con i TA in italiano e uno con i TA in inglese. Tale configurazione consente di utilizzare il corpus sia come corpus parallelo, sia come corpus comparabile. Inoltre, è stato effettuato l'allineamento testo-suono e l'allineamento dei TP con i rispettivi TA sulla base del contenuto (unità di informazione). Il corpus è annotato grammaticalmente, lemmatizzato e presenta anche l'annotazione di due tratti paralinguistici: le parole troncate e le disfluenze di pronuncia, secondo le medesime convenzioni applicate in EPIC. Il corpus è accessibile da un'apposita interfaccia web, sviluppata in collaborazione con il *Laboratorio de Lingüística Informática - Universidad Autónoma de Madrid* (LLI-UAM) e può essere studiato con gli strumenti della *Corpus Work Bench* (CWB-CQP), essendo stato pure indicizzato secondo uno standard compatibile con CWB. Ad ogni modo, anche in questo caso la cura nella scelta del formato in cui sono stati salvati i dati trascritti e registrati garantisce un buon grado di flessibilità nell'uso dei dati con diversi tipi di applicazioni.

Bisogna riconoscere che le dimensioni di entrambi i corpora, EPIC e DIRSI-C, sono probabilmente ancora troppo esigue per poter ottenere risultati generalizzabili o statisticamente significativi. Ciononostante, è stata finalmente creata una solida base metodologica e un primo standard da cui muovere i prossimi passi nell'avanzamento dei CIS (Bendazzoli & Sandrelli 2009) e, più in generale, nello sviluppo degli Studi sull'interpretazione. Possiamo affermare infatti che l'applicazione del *corpus-based approach* ha comportato un ampio abbraccio interdisciplinare, coinvolgendo non solo molte delle sottodiscipline che fanno parte dei *Translation Studies*, ma anche discipline "esterne" a cui si sta rivolgendo sempre più attenzione, in quella che alcuni percepiscono come una svolta sociologica o sociolinguistica degli *Interpreting Studies* (Pöchhacker 2006, 2008; Torresi 2009).

Nonostante siano già stati svolti diversi studi sui materiali raccolti in EPIC e DIRSI-C, per ragioni di spazio abbiamo potuto soltanto presentare brevemente tali esempi di analisi (§5.1). In questa sede si è ritenuto opportuno dedicare ampio spazio all'esposizione dell'intero percorso seguito nello svolgimento dei due progetti, con i quali ci auguriamo di aver contribuito concretamente all'avanzamento dei *Corpus-based Interpreting Studies* sul piano teorico, metodologico e soprattutto pratico. I due principali prodotti di queste ricerche, ovvero gli archivi multimediali e i rispettivi corpora EPIC e DIRSI-C, sono ora a disposizione della comunità scientifica, dei formatori e degli studenti, ai quali rivolgiamo un invito ad avvicinarsi ai CIS e alla metodologia interdisciplinare su cui si basano.

Indice delle Tabelle

Indice delle Figure

Bibliografia

Aarts, Bas & April, McMahon (eds.) (2006) *The Handbook of English Linguistics*. Malden: Blackwell.

Aarts, Jan; de Hoom, Pieter & Nelleke, Oostdijk (eds.) (1993) *English Language Corpora: Design, Analysis and Exploitation. Papers from the thirteenth international conference on English language research on computerized corpora*. Amsterdam/Atlanta: Rodopi.

Ahrens, Barbara (2004) Non-verbal phenomena in simultaneous interpreting. Causes and Functions. In Hansen, G. et al. (eds.), pp.227-237.

Ahrens, Barbara (2005) Prosodic phenomena in simultaneous interpreting: A conceptual approach and its practical application. *Intepreting* 7/1, pp. 51-76.

Albano Leoni, Federico (2005) Tre progetti per l'italiano parlato. In Burr, E. (ed.), pp. 153-161.

Alfieri, Gabriella & Stefania, Stefanelli (2005) Lessici dell'Italiano Radiofonico. In Burr, E. (ed.), pp. 397-412.

Almgren, Margareta; Barreña, Andoni; Ezeizabarrena, María José; Idiazabal, Itziar & Brian MacWhinney (eds.) (2001) *Research on Child Language Acquisition. Proceedings of the 8th Conference of the International Association for the Study of Child Language*. Somerville MA: Cascadilla.

Anderman, Gunilla & Margaret, Rogers (eds.) (2008) *Incorporating Corpora: The Linguist and the Translator.* Clevedon: Multilingual Matters. Series: Translating Europe.

Archer, Dawn; Rayson, Paul; Wilson, Andrew & Tony, McEnery (eds.) (2003) *Proceedings of the Corpus Linguistics 2003 conference. Lancaster University (UK), 28 - 31 March 2003*. Online: <http://ucrel.lancs.ac.uk/publications/cl2003/index.htm>.

Armstrong, Susan (1997) Corpus-based methods for NLP and translation studies. *Interpreting* 2/1-2, pp. 141-162.

Aston, Guy & Lou, Burnard (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University press.

Aston, Guy & Lou, Burnard (eds.) (2001) *Corpora in the Description and Teaching of English. Papers from the 5th ESSE conference*. Bologna: CLUEB.

Aston, Guy (1997) Small and large corpora in language learning. In Lewandowska-Tomaszczyk, Barbara & Patrick James, Melia (eds.), pp. 51-62.

Aston, Guy (ed.) (2001) *Learning with Corpora*. Bologna: CLUEB.

Aston, Guy; Bernardini, Silvia & Dominic, Stewart (eds.) (2004) *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins.

Atkinson, J. Maxwell & John Heritage (1999) Jefferson's transcript notation. In Jarowski, A. & N., Coupland (eds.), pp. 158-166.

Austin, John (1976) *How to do things with words: the William James lectures delivered at Harvard University in 1955* [Urmson J. O. & Marina Sbisà (eds.)] London: Oxford University Press.

Azzaro, Gabriele & Margherita Ulrych (a cura di) (1999) *Atti del XVIII Convegno AIA. Genova, 30 settembre - 2 ottobre 1996. Anglistica e…: metodi e percorsi comparatistici nelle lingue, culture e letterature di origine europea. Volume II. Transiti Linguistici e Culturali*. Trieste: E.U.T.

Baker, M.; Francis, G. & E. Tognini-Bonelli (eds.) (1993) *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins.

Baker, Mona (1993) Corpus Linguistics and Translation Studies: implications and applications. In Baker, M. et al. (eds.), pp. 233-250.

Baker, Mona (1995) Corpora in Translation Studies. An overview and suggestions for future research. *Target* 7/2, pp. 223-243.

Baker, Mona (1996) Corpus-based Translation Studies. The challenges that lie ahead. In Somers, H. (ed.), pp. 175-186.

Baker, Mona (1999) The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4/2, pp. 281-298.

Baldry, Anthony & J. Paul, Thibault (2001) Towards Multimodal corpora. In Aston, G. & L. Burnard (eds.), pp. 277-305.

Baldry, Anthony & J. Paul, Thibault (2005) *Multimodal Transcription and Ttext Analysis: A Multimedia Toolkit and Coursebook*. London/Oakville: Equinox.

Bally, Charles (1971) *Linguistica generale e linguistica francese*. Con introduzione e appendice di Cesare Segre; traduzione di Giovanni Caravaggi. Milano: Il Saggiatore.

Barlow, Michael (2001/2003) *ParaConc: A Concordancer for Parallel Texts (D ra f t 3 / 0 3)*. Online: <http://www.athel.com/paraconc.pdf>.

Baroni, Marco & Silvia Bernardini (eds.) (2006). *Wacky! Working papers on the Web as Corpus*. Bologna: Gedit.

Baroni, Marco; Bernardini, Silvia; Comastri, Federica; Piccioni, Lorenzo; Volpi, Alessandra; Aston, Guy & Marco, Mazzoleni (2004) Introducing the La Repubblica Corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In Lino, Maria Teresa et al. (eds.), Vol. 5, pp. 1771-1774. Online: <http://dev.sslmit.unibo.it/corpora/downloads/rep_lrec_2004.pdf>.

Baroni, Marco; Bernardini, Silvia; Ferraresi Adriano & Eros, Zanchetta (2009) The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation* 43/3, pp. 209-226.

Baroni, Marco; Kilgarriff, Adam; Pomikálek, Jan & Pavel, Rychlý (2006) WebBootCaT: Instant domain-specific corpora to support human translators. In *Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation.* Oslo: The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway), pp. 247-252. 2006.

Baugh, J. & J. Sherzer (eds.) (1984) *Language in Use: Readings in Sociolinguistics*. Englewood Cliffs, NJ: Prentice Hall.

Bayne, Kristofer (2005) Using the 'poster session' format in L2 contexts. In Ross, P. et al. (eds.), pp. 87-103. Online: <http://jalt.org/pansig/2005/HTML/Bayne.htm>.

Bazzanella, Carla (1994) *Le facce del parlare: un approccio pragmatico all'italiano parlato*. Scandicci: La Nuova Italia.

Beeby, Allison; Rodríguez Inés, Patricia & Pilar, Sánchez-Gijón (eds.) (2009) *Corpus Use and Translating. Corpus use for learning to translate and learning corpus use to translate*. Amsterdam/Philadelphia: John Benjamins.

Bendazzoli, Claudio & Annalisa, Sandrelli (2005/2007) An approach to corpus-based interpreting studies: developing EPIC (European Parliament Interpreting Corpus). In

Gerzymisch-Arbogast H. & S. Nauert (eds.). Online: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html>.

Bendazzoli, Claudio & Annalisa, Sandrelli (2009) Corpus-based Interpreting Studies: Early work and future prospects. *Tradumatica 7. L'aplicació dels corpus linguistics a la traducció*. Online: <http://webs2002.uab.es/tradumatica/revista/num7/articles/08/08art.htm>.

Bendazzoli, Claudio (2010) The European Parliament as a source of material for research into simultaneous interpreting: advantages and limitations. In Zybatow, N. L. (ed.), pp. 51-68.

Bendazzoli, Claudio; Monti, Cristina; Sandrelli, Annalisa; Russo, Mariachiara; Baroni, Marco; Bernardini, Silvia; Mack, Gabriele; Ballardini, Elio & Peter, Mead (2004) Towards the creation of an electronic corpus to study directionality in simultaneous interpreting. In N. Oostdijk, Kristoffersen, G. & G. Sampson (eds.), pp. 33-39.

Bendazzoli, Claudio; Sandrelli, Annalisa & Mariachiara, Russo (in corso di stampa 2011) Disfluencies in simultaneous interpreting: a corpus-based analysis. In Kruger, Alet; Walmach, Kim & Jeremy, Munday (eds.) *Corpus-based Translation Studies Research and Applications.* London/New York: Continuum.

Bernardini, Silvia & Sara, Castagnoli (2008) Corpora for translator education and translation practice. In Yuste, E. (ed.), pp. 39-55.

Bernardini, Silvia (2000) *Competence, Capacity, Corpora: A Study in Corpus-aided Language Learning.* Bologna: CLUEB.

Berruto, Gaetano (1997) *Fondamenti di Sociolinguistica.* Roma: Laterza.

Bersani Berselli, Gabriele (2004) Linguistica ed interpretazione: la conferenza come genere testuale. In Bersani Berselli, G., Mack, G. & D. Zorzi (eds.), pp. 35-71.

Bersani Berselli, Gabriele; Mack, Gabriele & Daniela, Zorzi (eds.) (2004) *Linguistica e Interpretazione*. Bologna: CLUEB.

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

Bilbow, T. Grahame (2007) Speaking and not speaking across cultures. In Garzone, Giuliana & Cornelia, Ilie (eds.), pp. 229-224.

Bird, Steven; Buneman, Peter & Mark, Liberman (eds.) (2001) *Proceedings of the IRCS Workshop On Linguistic Databases. 11-13 December 2001, University of Pennsylvania, Philadelphia, USA*. Online: <http://www.ldc.upenn.edu/annotation/database/proceedings.html>.

Blanche-Benveniste, Claire (2005) *Estudios lingüísticos sobre la relación entre oralidad y escritura*. Sevilla: Gedisa.

Boersma, Paul & David, Weenink (2001) Praat, a system for doing phonetics by computer. *Glot International* 5/9-10, pp. 341-345.

Bondi, Marina; Gavioli, Laura & Marc, Silver (eds.) (2004) *Academic Discourse, Genre and Small Corpora.* Roma: Officina.

Bortolini, Umberta & Elena, Pizzuto (a cura di) (1997) *Il progetto CHILDES-Italia: contributi di ricerca sulla lingua italiana.* Tirrenia, Pisa: Edizioni del Cerro.

Bowker, Lynne & Jennifer, Pearson (2002) *Working with Specialized Language. A practical guide to using corpora.* London/New York: Routledge.

Bowker, Lynne (1998) Using specialized monolingual native-language corpora as a translation resource: a pilot study. *Meta* 43/4, pp. 631-651. Online: <http://id.erudit.org/iderudit/002134ar>.

Bowker, Lynne (2002) *Computer-Aided Translation Technology: A Practical Introduction.* Ottawa: Ottawa University Press.

Braun, Sabine (2006a) ELISA - A pedagogically enriched corpus for language learning purposes. In Braun, S. et al. (eds.), pp. 25-47.

Braun, Sabine; Kohn, Kurt & Joybrato, Mukherjee (eds.) (2006) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang.

Breiteneder, Angelika; Klimpfinger, Theresa; Majewski, Stefan; Pitzl, Marie-Luise (2009) The Vienna-Oxford International Corpus of English (VOICE) - A linguistic resource for exploring English as a lingua franca. *ÖGAI-Journal* 28/1, pp. 21-26.

Brown, Gillian & George, Yule (1986) *Analisi del discorso* (Traduzione in italiano di Giuliano Bernini). Bologna: Il Mulino.

Burnard, Lou & C. Michael, Sperberg McQueen (eds.) (1994) *Guidelines for Electronic Text Encoding and Interchange*. Chicago/Oxford: ACH, ACL, ALLC.

Burnard, Lou (1995) The Text Encoding Initiative: an overview. In Leech, G. et al. (eds.), pp. 69-81.

Burr, Elisabeth (a cura di) (2005) *Tradizione e innovazione. Il parlato: teoria, corpora, linguistica dei corpora. Atti del VI Convegno Internazionale della SILFI, Duisburg 28.06.-02.07.2000* (Quaderni della Rassegna 43). Firenze: Franco Cesati Editore.

Calzada Pérez, María & Luz, Saturnino (2006) ECPC - Technology as a Tool to Study the (Linguistic) Functioning of National and Trans-National European Parliaments. *International Journal of Technology, Knowledge and Society* 2/5, pp. 53-61.

Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odjik, Jan; Piperidis, Stelios & Daniel, Tapias (eds.) (2008) *Proceedings of the Sixth International Language Resources and Evaluation (LREC '08)*. ELRA. Online: <http://www.lrec-conf.org/proceedings/lrec2008/>.

Campillos, Leonardo; Gozalo, Paula; Guirao, José María & Antonio, Moreno Sandoval (2007) Exploiting a spoken corpus in language teaching/learning: An advanced web-based tool. *Proceedings from Corpus Linguistics Conference Series.CL2007, University of Birmingham, UK, July 27-30 2007*. Online: <http://www.corpus.bham.ac.uk/corplingproceedings07/paper/99_Paper.pdf>.

Campoy, Mari Carmen & María José, Luzón (eds.) (2007) *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang.

Cantos Gómez, Pascual & Aquilino, Sánchez Pérez (eds.) (2009) *A Survey on Corpus-based Research / Panorama de investigaciones basadas en corpus*. Murcia: AELINCO (Asociación Española de Lingüística de Corpus).

Carabelli, Angela (2003) A brief overview of IRIS – the Interpreters' Research Information System. In de Manuel Jerez, Jesús (coord.), pp. 113-139.

Carlson, Lynn; Marcu, Daniel & Mary Ellen, Okurowski (2001) Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*. Online: <http://www.aclweb.org/anthology-new/W/W01/W01-1605.pdf>.

Carreras, Xavier; Chao, Isaac; Padró, Lluís & Muntsa, Padró (2004) Freeling: an open-source suite of language analyzers. In Lino, Maria Teresa et al. (eds.), pp. 239-242.

Castagnoli, Sara (2008) *Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation*. Tesi di dottorato, Dipartimento di Linguistica, Università di Pisa.

Castagnoli, Sara (2009) A new approach to the analysis of explicitation in Translation: Multiple (Learner) Translation Corpora. *International Journal of Translation* 21/1, pp. 89-105.

Castello, E. (2004) Calcolo della densità lessicale e dell'intricatezza grammaticale di corpora linguistici. In Taylor Torsello, C. et al. (eds.), pp. 131-151.

Cavagnoli, Stefania; Di Giovanni, Elena & Raffaela, Merlini (eds.) (2009) *La ricerca nella comunicazione interlinguistica: modelli teorici e metodologici*. Milano: Franco Angeli.

Cecot, Michela (2001) Pauses in simultaneous interpretation: a contrastive analysis of professional interpreters' perfomances. *The Interpreters' Newsletter* 11, pp. 63-85. Online: <http://www.openstarts.units.it/dspace/bitstream/10077/2448/1/04.pdf>.

Cencini, Marco & Guy, Aston (2002) Resurrecting the corp(us/se): Towards an encoding standard for interpreting data. In Garzone, G. & M. Viezzi (eds.), pp. 47-62.

Cencini, Marco (2000) Il Television Interpreting Corpus (TIC). *Proposta di codifica conforme alle norme TEI per trascrizioni di eventi di interpretazione in televisione.* Tesi di laurea non pubblicata, Università di Bologna, Sede di Forlì, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Cencini, Marco (2002) On the importance of an encoding standard for corpus-based interpreting studies. *inTRAlinea* Special Issue: CULT2K. Online: <http://www.intralinea.it/specials/cult2k/eng_open.php?id=P107>.

Chafe, Wallace (1995) Adequacy, user-friendlyness, and practicality in transcribing. In Leech, Geoffrey. et al. (eds.), pp. 54-61.

Chang, Chia-Chien & L. Diane, Schallert (2007) The impact of directionality on Chinese/English simultaneous interpreting . *Interpreting* 9/2, 2007, pp. 137-176.

Chiari, Isabella (2006) Slips and errors in spoken data transcription. In *Proceedings of the LREC 2006 Conference, Genova, Magazzini del Cotone 24-26 May 2006*. Genova: ELRA. Online: <http://hnk.ffzg.hr/bibl/lrec2006/>.

Chiaro, Delia; Heiss, Christine & Chiara, Bucaria (eds.). *Between Text and Image: Updating Research in Screen Translation.* Amsterdam/Philadelphia: John Benjamins.

Christ, Oli (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System, *COMPLEX '94, Budapest*. Online: <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/#Papers>.

Cokely, Dennis (1992) *Interpretation: A Sociolinguistic Model*. Burtonsville: Linstok Press.

Collados Aís, Ángela; Fernández Sánchez, María Manuela & Daniel, Gile (eds.) (2003) *La evaluación de la calidad en interpretación: Investigación*. Granada: Comares.

Comastri, Federica (2002) Un esperimento nella creazione di un testo elettronico parallelo. Codifica e allineamento di *A Brief History of Time* di Stephen Hawking. *inTRAlinea* 5. Online: <http://www.intralinea.it/volumes/ita_more.php?id=135_0_2_0_C>.

Cook, Guy (1995) Theoretical issues: transcribing the untranscibable. In Leech, Geoffrey. et al. (eds.), pp. 35-53.

Cook, Guy (2010) *Translation in Language Teaching: An Argument for Reassessment*. Oxford: Oxford University Press.

Corpas Pastor, Gloria (2008) *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frakfurt am Main: Peter Lang.

Cosmai, Domenico (2003) *Tradurre per l'Unione europea: problematiche e strategie operative*. Milano: Hoepli.

Coveri, Lorenzo (a cura di) (1984) *Linguistica Testuale. Atti del XV Cogresso Internazionale di Studi. Genova – Santa Margherita Ligure, 8-10 Maggio 1981*. Roma: Bulzoni.

Crawford Camiciottoli, Belinda (2004) Non-verbal communication in intercultural lectures. In Bondi, Marina et al. (eds.), pp. 35-51.

Cresti, Emanuela & Massimo, Moneglia (eds.) (2005) *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins.

Cresti, Emanuela (1995) Speech act units and informational units. In Fava, E. (ed.), pp. 89-107.

Cresti, Emanuela (2000a) *Corpus di italiano parlato. Volume I: Introduzione*. Firenze: presso l'Accademia della Crusca.

Cresti, Emanuela (ed.) (2000b) *Corpus di italiano parlato. Volume II: Campioni*. Firenze: presso l'Accademia della Crusca.

Cresti, Emanuela; Panunzi Alessandro & Antonietta, Scarano (2005) The Italian corpus. In Cresti, Emanuela & Massimo, Moneglia (eds.), pp. 71-110.

D'hondt, Sigurd; Östman, Jan-Ola & Jef, Verschueren (eds.) (2009) *The Pragmatics of Interaction.* Amsterdam/Philadelphia: John Benjamins.

Danielsson, Pernilla (2004) Programming: Simple Perl programming for corpus work. In Sinclair, John (ed.), pp. 225–246.

Dayrell, Carmen (2005) *Investigating lexical patterning in translated Brazilian Portuguese: a corpus-based study*. Unpublished PhD Thesis. Manchester: The University of Manchester.

Dayrell, Carmen (2007) A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics* 12/3, pp. 375–414.

de Manuel Jerez, Jesús (2003a) El canal EbS en la mejora de la calidad de la interpretación: perfiles profesionales de especialidad en el itinerario de interpretación. In Collados Aís, A. et al. (eds.) (2003), pp. 207-218.

de Manuel Jerez, Jesús (2003b) Nuevas tecnologías y selección de contenidos: la base de datos *Marius*. In de Manuel, J. (ed.) (2003), pp. 21-65.

de Manuel Jerez, Jesús (ed.) (2003) *Nuevas tecnologías y formación de intérpretes*. Granada: Comares.

De Martino, Marco (2006/2007) *Case study dell'interpretazione simultanea di un convegno medico*. Tesi di laurea non pubblicata, Università di Bologna, Sede di Forlì, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

De Mauro, Tullio, Mancini, Federico, Vedovelli, Massimo & Miriam, Voghera (1993) *Lessico di frequenza dell'italiano parlato*. Etas: Milano.

Di Guida, F. (2001) *Analisi del relay in interpretazione simultanea sulla base di dati empirico-descrittivi*. Tesi di laurea non pubblicata, Università di Trieste, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Di Kearns, John (ed.) *Translator and Interpreter Training: Issues, Methods and Debates*. London/New York: Continuum.

Dimitrova, Englund Birgitta & Kenneth, Hyltenstam (eds.) (2000) *Language Processing and Simultaneous Interpreting: Interdisciplinary Perspectives*. Amsterdam/Philadelphia: John Benjamins.

Diriker, Ebru (2004) *De-/Re-Contextualizing Conference Interpreting. Interpreters in the Ivory Tower?* Amsterdam/Philadelphia: John Benjamins.

Dollerup, Cay & Annette, Lindegaard (eds.) (1992) *Teaching Translation and Interpreting. Training, Talent and Experience. Papers from the First Language International Conference. Elsinore, Denmark 31 May – 2 June 1991*. Amsterdam/Philadelphia: John Benjamins.

Dollerup, Cay & Annette, Lindegaard (eds.) (1994) *Teaching Translation and Interpreting 2. Insights, Aims and Visions. Papers from the Second Language International Conference. Elsinore, Denmark 1993*. Amsterdam/Philadelphia: John Benjamins.

Dollerup, Cay & Leo, Ceelen (1996) *A Corpus of Consecutive Interpreting. Comprising Danish, Dutch, English, French, German and Italian*. Copenhagen: Centre for Translation Studies and Lexicogrphy, University of Copenhagen.

Donato, Valentina (2003) Strategies adopted by student interpreters in SI: a comparison between the English-Italian and the German-Italian language-pairs. *The Interpreters' Newsletter* 12, pp. 101-134.

Du Bois, W. John & Robert, Englebretson (2004) *Santa Barbara corpus of spoken American English, Part 3*. Philadelphia: Linguistic Data Consortium. ISBN 1-58563-308-9.

Du Bois, W. John & Robert, Englebretson (2005) *Santa Barbara corpus of spoken American English, Part 4*. Philadelphia: Linguistic Data Consortium. ISBN: 158563-348-8.

Du Bois, W. John (2006a) SoundWriter 2.0 Manual. Online: <http://www.linguistics.ucsb.edu/projects/transcription/soundwriter.pdf>.

Du Bois, W. John (2006b) VoiceWalker. A discourse transcription utility. Online: <http://www.linguistics.ucsb.edu/projects/transcription/voicewalker.pdf>.

Du Bois, W. John; Chafe, L. Wallace; Meyer, Charles & A. Sandra, Thompson (2000) *Santa Barbara corpus of spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium. ISBN 1-58563-164-7.

Du Bois, W. John; Chafe, L. Wallace; Meyer, Charles; Thompson, A. Sandra & Nii, Martey (2003) *Santa Barbara corpus of spoken American English, Part 2*. Philadelphia: Linguistic Data Consortium. ISBN 1-58563-272-4.

Du Bois, W. John; Schuetze-Coburn, Stephan; Cumming, Susanna & Danae, Paolino (1993) Outline of discourse transcription. In Edwards, J.& M. Lampert (eds.), pp. 45-89.

Duranti, Alessandro (1997) *Linguistic Anthropology*. New York: Cambridge University Press.

Edwards, A. Jane & Martin, D. Lampert (eds.) (1993) *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Edwards, A. Jane (1993a) Principles and contrasting systems of discourse transcription. In Edwards, J.& M. Lampert (eds.), pp. 3-32.

Edwards, A. Jane (1993b) Survey of electronic corpora and related resources for language researchers. In Edwards, J.& M. Lampert (eds.), pp. 263-306.

Edwards, A. Jane (1995) Principles and alternative systems in the transcription, coding and murk-up of spoken discourse. In Leech, Geoffrey et al. (eds.), pp. 19-34.

Ehlich, Konrad (1993) HIAT: a transcription system for Discourse data. In Edwards, A. Jane & Martin, D. Lampert (eds.), pp. 123-148.

Facchinetti, Roberta & Matti, Rissanen (eds.) (2006) *Corpus-based Studies of Diachronic English*. Bern: Peter Lang.

Facchinetti, Roberta (ed.) (2007) *Corpus Linguistics 25 Years On*. Amsterdam/New York: Rodopi.

Falbo, Caterina (2005) La transcription: une tache paradoxale. *The Interpreters' Newsletter* 13, pp. 25-38. Online: <http://www.openstarts.units.it/dspace/bitstream/10077/2468/1/03.pdf>.

Falbo, Caterina; Russo, Mariachiara & Francesco, Straniero Sergio (a cura di) (1999) *Interpretazione simultanea e consecutiva. Problemi teorici e metodologie didattiche*. Milano: HOEPLI.

Ferraresi, Adriano (2009) Google and Beyond: Web-As-Corpus Methodologies for Translators. *Tradumatica 7. L'aplicació dels corpus linguistics a la traducció*. Online: <http://webs2002.uab.es/tradumatica/revista/num7/articles/04/04art.htm >.

Firenze, Laura (2002) *Aspetti interazionali del discorso monologico: la conferenza*. Tesi di laurea non pubblicata, Università di Bologna – Sede di Forlì, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Firenze, Laura (2004) Interpretare gli aspetti interpersonali della conferenza. In Bersani Berselli, G. et al. (eds.), pp. 147-167.

Fumagalli, D. (1999/2000) *Alla ricerca dell'interpretese. Uno studio sull'interpretazione consecutiva attraverso la corpus linguistics*. Tesi di laurea non pubblicata, Università di Trieste, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Gagliardi, Cesare (a cura di) (1997) *Imparare ad imparare nei centri linguistici multimediali*. Pescara: Libreria dell'università.

Galatolo Renata & Gabriele Pallotti (a cura di) (1999) *La conversazione. Un'introduzione allo studio dell'interazione verbale*. Milano: Raffaello Cortina Editore.

Galli, Cristina (1988/1989) L'interpretazione simultanea nei congressi di medicina: un contributo sperimentale. Tesi di laurea in interpretazione. Trieste: Università degli studi di Trieste, Scuola superiore di lingue moderne per interpreti e traduttori.

Galli, Cristina (1990) Simultaneous interpretation in medical conferences: a case-study. In Gran, Laura & Christopher, Taylor (eds.), pp. 61-82.

Garrote Salazar, Marta (2008) *CHIEDE. Corpus de Habla Infantil Espontánea del Español*. Tesi di dottorato, Universidad Autónoma de Madrid.

Garside, Roger; Fligelstone, Steve & Simon, Botley (1997a) Discourse annotation: anaphoric relations in corpora. In Garside, Roger et al. (eds.) (1997), pp. 66-84.

Garside, Roger; Leech, Geoffrey & Anthony, Mc Enery (eds.) (1997) *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London/New York: Longman.

Garwood, J. Christopher (2002) Autonomy of the interpreted text. In Garzone, G. & M. Viezzi (eds.), pp. 267-276.

Garzone, Giuliana & Cornelia, Ilie (eds.) (2007) *The Use of English in Institutional and Business Settings. An intercultural perspective*. Bern: Peter Lang.

Garzone, Giuliana & Maurizio, Viezzi (2001) *Comunicazione specialistica e interpretazione di conferenza*. Trieste: Edizioni Università di Trieste.

Garzone, Giuliana & Maurizio, Viezzi (eds.) (2002) *Interpreting in the 21st Century. Challenges and Opportunities. Selected papers from the first Forlì Conference on Interpreting Studies, 9-11 November 2000*. Amsterdam/Philadelphia: John Benjamins.

Garzone, Giuliana; Mead, Peter & Maurizio, Viezzi (eds.) (2002) *Perspectives on Interpreting*. Bologna: CLUEB.

Gavioli, Laura & Gillian, Mansfield (eds.) (1990) *The PIXI Corpora: Bookshop Encounters in English and Italian*. Bologna: CLUEB.

Gavioli, Laura (1999) Alcuni meccanismi di base dell'analisi della conversazione. In Galatolo, R. & G. Pallotti (a cura di), pp. 43-65.

Gerver, David & H. Wallace, Sinaiko (eds.) (1978) *Language Interpretation and Communication*. New York: Plenum Press.

Gerzymisch-Arbogast, Heidrun & Sandra, Nauert (eds.) (2005/2007) *Proceedings of the Marie Curie Euroconferences. MuTra: Challenges of Multidimensional Translation — Saarbrücken 2-6 May 2005*. Online: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html>.

Gerzymisch-Arbogast, Heidrun; Hajicová, Eva; Sgall, Peter; Jettmarová, Zuzana; Rothkegel, Annely & Dorothee, Rothfuß-Bastian (eds) (2003) *Textologie und Translation*. Tübingen: Gunter Narr.

Ghadessy, Mohsen; Henry, Alex & L. Robert, Roseberry (eds.) (2001) *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam/Philadelphia : John Benjamins.

Gile, Daniel (1994) Opening up Interpretation Studies. In Snell-Hornby, M. et al. (eds.), pp. 149-158.

Gile, Daniel (1998) Observational studies and experimental studies in the investigation of conference interpreting. *Target* 10/1, pp. 69-93.

Godfrey, J. John; Edward C. & McDaniel, Holliman Jane (1992) SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP-92 International Conference on Acoustics, Speech, and Signal Processing, 1992*. Vol. 1, pp. 517-520.

Godijns, R. & M. Hinderdael (eds.) (2005) *Directionality in Interpreting. The 'Retour' or the Native?* Gent: Communication and Cognition.

Goedertier, W.; Goddijn, S. & J.P. Martens (2000) Orthographic Transcription of the Spoken Dutch Corpus. In Gravilidou, M. et al. (eds.), pp. 909-914.

Goffman, Erving (1981) *Forms of Talk*. Oxford: Blackwell.

González Rodríguez, Manuel & Carmen Paz, Suárez Araujo (eds.) (2002) *LREC 2002. Proceedings of the Third International Conference on Language Resources and Evaluation, 29th, 30th & 31st May 2002, Las Palmas de Gran Canaria*. Paris: ELRA.

Gran, Laura & Alessandra, Riccardi (a cura di) (1997) *Nuovi Orientamenti negli Studi sull'Interpretazione*. Padova: CLEUP.

Gran, Laura & Christopher, Taylor (eds.) (1990) *Aspects of Applied and Experimental Research on Conference Interpretation: Round Table on Interpretation Research, November 16, 1989*. Udine: Campanotto.

Graupera, M. (2009) *La interpretación simultánea en el Parlamento Europeo: estudio contrastivo del uso del perfetto composto en italiano, español e inglés*. Tesi di laurea non pubblicata, Facoltà di Interpretariato e Traduzione, LUSPIO Roma.

Gravilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S. & G. Stainhaouer (eds.) (2000) *LREC 2000. Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 31st May-2nd June 2000*. Paris: ELRA.

Greenbaum, Sidney (ed.) (1996) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon press.

Guirao, José María & Antonio, Moreno Sandoval (2004) A "toolbox" for tagging the Spanish C-ORAL-ROM corpus. In *Proceedings of the IV International Conference on Language Resources and Evaluation (LREC2004)*. Lisbon: ELRA.

Halliday, M.A.K. (1992) *Lingua parlata e lingua scritta*. Firenze: La Nuova Italia.

Halverson, Sandra (1998) Translation Studies and representative corpora: Establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta* 43/4, pp. 494-514/1-22. Online: <http://www.erudit.org/revue/meta/1998/v43/n4/003000ar.pdf>.

Hansen, Gyde; Chesterman, Andrew & Heidrun, Gerzymisch-Arbogast (eds.) (2008) *Efforts and Models in Interpreting and Translation Research. A tribute to Daniel Gile*. Amsterdam/Philadelphia: John Benjamins.

Hansen, Gyde; Malmkjaer, Kirsten & Daniel Gile (eds.) *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*. Amsterdam/Philadelphia: John Benjamins.

Hayashi, Reiko (1996) *Cognition, Empathy, and Interaction: Floor Management of English and Japanese Conversation*. Norwood NJ: Ablex.

Heiss, Christine & Marcello, Soffritti (2008). Forlixt 1 — The Forlì Corpus of Screen Translation: Exploring Microstructures. In. Chiaro, D. et al. (eds.). pp. 51-62.

Heritage, John (1995) Conversation Analysis: Methodological aspects. In Quasthoff, U.M. (ed.), pp. 391-416.

Heritage, John (1997) Conversation analysis and institutional talk: analyzing data. In Silverman, D. (ed.), pp. 161-182.

Hidalgo, Antonio & Xose A., Padilla (2006) Bases para el análisis de las unidades menores del discurso oral: los subactos. *ORALIA* 9, pp. 109-143.

Hidalgo, Encarnación; Quereda, Luis & Juan, Santana (eds.) (2007) *Corpora in the Foreign Language Classroom: Selected Papers from the 6th International Conference on Teaching and Language Corpora (TaLC 6), University of Granada, Spain, 4-7 July, 2004*. Amsterdam/New York: Rodopi.

Hofland, Knut & Stig, Johansson (1998) The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In Johansson, S. & S. Oksefjell (eds.), pp. 87-100.

Hofland, Knut (2003) A web-based concordance system for spoken language corpora. In Archer, Dawn et al. (eds.), pp. 330-331. Online: <http://ucrel.lancs.ac.uk/publications/cl2003/papers/hofland_abstract.pdf>.

Hurtado Albir, A. (1994/1995) Modalidades y tipos de traducción. *Vasos Comunicantes 4*, pp. 19-27. Online: <http://www.acett.org/numero.asp?numero=4>.

Hurtado Albir, A. (1996) La traduction: classification et elements d'analyse. *Meta* 41/1, pp. 366-377.

Hutchby, Ian & Robin, Wooffitt (1999) *Conversation Analysis. Principles, Practices and Applications*. Cambridge: Polity Press.

Hymes, Dell (1980) *Fondamenti di sociolinguistica. Un approccio etnografico*. Traduzione di Filippo Beghelli. Revisione di Gaetano Berruto. Bologna: Zanichelli.

Ibrahim, Noraimi (2009) Parliamentary Interpreting in Malaysia: A Case Study. *Meta* 54/2, pp. 357-369. Online: <http://id.erudit.org/iderudit/037686ar>.

Ivanova, Adelina (2000) Designing the retrospective study: Research and methodological issues. In Tirkkonen-Condit, S. & R. Jääskeläinen (eds.), pp. 31-52.

Izre'el, Shlomo; Hary, Benjamin & Giora, Rahav (2001) Designing CoSIH: the corpus of spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6/2, pp. 171–197.

Jantunen, Jarmo (2008) *Il corpus della lingua finlandese: collaborazione tra tra le Università di Oulu e Forlì*. Comunicazione presentata in occasione della Giornata di Studi di Lingua, Letteratura e Cultura Finlandese, 13 maggio 2008, Dipartimento SITLeC, Università di Bologna, sede di Forlì.

Jarowski, Adam & Nikolas, Coupland (eds.) (1999) *The Discourse Reader*. London/New York: Routledge.

Jiménez Ivars, Amparo (1999) *La traducción a la vista. Un análisis descriptivo*. Tesi di dottorato, Universitat Jaume I de Castellón (Spagna), Online: <http://www.tdr.cesca.es/>.

Jiménez Ivars, Amparo (2002) Variedades de interpretación: modalidades y tipos. *Hermeneus IV*, pp. 95-114.

Johansson, Stig & Signe, Oksefjell (eds.) (1998) *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam/Atlanta GA: Rodopi.

Johansson, Stig (1995) The approach of the Text Encoding Initiative to the encoding of spoken discourse. In Leech, Geoffrey et al. (eds.), pp. 82-98.

Johansson, Stig (1998) On the role of corpora in cross-linguistic research. In Johansson S. & S. Oksefjell (eds.), pp. 3-24.

Johansson, Stig (2007) *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam/Philadelphia: John Benjamins.

Johnstone, B. (2000) *Qualitative Methods in Sociolinguistics*. New York/Oxford: Oxford University Press.

Jurafsky, Daniel & James H. Martin (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River NJ: Prentice Hall.

Kahrel, Peter; Barnett, Ruthanna & Geoffrey, Leech (1997) Towards cross-linguistic standards or guidelines for the annotation of corpora. In Garside, Roger et al. (eds.), pp. 231-242.

Kalina, Sylvia (1994) Analyzing interpreters' performance: methods and problems. In C. Dollerup & A. Lindegaard (eds.), pp. 225-232.

Kalina, Sylvia (1998) *Strategische Prozesse beim Dolmetschen. Theoretische Grundlagen, empirische Fallstudien, didaktische Konsequenzen*. Tübingen: Gunther Narr.

Kalina, Sylvia (2000) Interpreting competences as a basis and a goal for teaching. *The Interpreters' Newsletter* 10, pp. 3-32.

Kalina, Sylvia (2005) Quality in the interpreting process: what can be measured and how? In Godijns, R. & M. Hinderdael (eds.), pp. 27-46.

Kaunzner, Ulrike & Frederik, Gianni (1997) Il progetto Audio Lingua: miglioramento della comprensione uditiva e della espressione orale di una lingua straniera. In Gagliardi, C. (a cura di), pp. 37-78.

Kaunzner, Ulrike (1997) Sound Perception Training and Foreign Language Learning. In Stame, S. (ed.), pp. 110-121.

Kawaguchi, Yuji; Zaima, Susumu & Toshihiro, Takagaki (eds.) (2006) *Spoken Language Corpus and Linguistic Informatics*. Amsterdam/Philadelphia: John Benjamins.

Kellett Bidoli, Cynthia Jane (2004) Intercultural features of English-to-Italian sign language conference interpretation: a preliminary study for Multimodal Corpus Analsys. *Textus* 17, pp. 127-142.

Kellett Bidoli, Cynthia Jane (2007) The Linguistics conference setting: a comparative analysis of intercultural disparities during Elnglish to Italian sign language interpretation. In Garzone, Giuliana & Cornelia, Ilie (eds.), pp. 331-349.

Kenny, Dorothy (2001) *Lexis and Creativity in Translation. A corpus-based study*. Manchester: St. Jerome.

King, P. & D. Woolls, (1996) Creating and using a multilingual parallel concordancer. *Translation and Meaning* 4, pp. 459-466.

Klaudy, Kinga & Janos, Kohn (eds.) (1997) *Transferre Necesse Est. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting: 5-7 September 1996, Budapest, Hungary*. Budapest: Scholastica.

Knowles, Gerry (1993) The machine-readable Spoken English Corpus. In Aarts, Jan et al. (eds.), pp.107-119.

Kohnen, Thomas (2000) Corpora and speech acts: The study of performatives. In Mair, Christian & Marianne, Hundt (eds.), pp. 177-186.

Kolb, Waltraud & Franz, Pöchhacker (2008) Stories Retold: Interpreting in Asylum Appeal Hearings. In Russell D. & S. Hale (eds.), pp. 26-50.

Koller, Werner (1995) The concept of equivalence and the object of Translation Studies. *Target* 7/2, pp. 191-222.

Kruger, Alet (2004) Editorial: Corpus-based translation research comes to Africa. *Language Matters. Studies in the Languages of Africa* 35/1, Special Issue: Corpus-based Translation Studies: Research and Applications, pp. 1-5.

Kruger, Alet (2008) Translation, self-translation and apartheid-imposed conflict. *To be published in Language and Politics 2008*. Online: <http://www.multilingua.co.za/pdfs/Kruger_2008_Language_and_Politics.pdf>.

Kurz, Ingrid (2002) Physiological stress responses during media and conference interpreting. In Garzone, G. & M. Viezzi (eds.), pp. 195-202.

Kurz, Ingrid (2003) Physiological stress during simultaneous interpreting: a comparison of experts and novices. *The Interpreters' Newsletter* 12, pp. 51-67.

Labov, W. (1966) *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.

Labov, W. (1984) Field methods of the project on linguistic change and variation. In Baugh, J. & J. Sherzer (eds.), pp. 28-66.

Lambert, Sylvie & Barbara, Moser Mercer (eds.) (1994) *Bridging the gap: empirical research in simultaneous interpretation.* Amsterdam/Philadelphia: John Benjamins.

Lambert, Sylvie (1992) Shadowing. *The Interpreters' Newsletter* 4, pp. 15-24.

Laviosa, Sara (1998a) The Corpus-based Approach: A new Paradigm in Translation Studies. *Meta* 43/4, pp. 474-479. Online: <http://id.erudit.org/iderudit/003424ar>.

Laviosa, Sara (1998b) Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43/4, pp. 557-570.

Laviosa, Sara (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam/New York: Rodopi.

Laviosa, Sara (2004) Corpus-based translation studies: Where does it come from? Where is it going? *Language Matters. Studies in the Languages of Africa* 35/1, Special Issue: Corpus-based Translation Studies: Research and Applications, pp. 6-27.

Lawson, A. (2001) Collecting, aligning and analysing parallel corpora. In Ghadessy, M. et al. (eds.), pp. 279-309.

Lederer, Marianne (1981) *La traduction simultanée – Expérience et théorie.* Paris: Minard Lettres Modernes.

Lee, Y.W. David (2008) ICAME Conferences and Proceedings. Online: <http://icame.uib.no/ICAME_Proceedings_Proceedings.pdf>.

Leech, Geoffrey & Elizabeth Eyes (1997) Syntactic annotation: treebanks. In Garside, Roger et al. (eds.), pp. 34-52.

Leech, Geoffrey (1997a) Introducing corpus annotation. In Garside, Roger et al. (eds.) (1997), pp. 1-18.

Leech, Geoffrey (1997b) Grammatical tagging. In Garside, Roger et al. (eds.) (1997), pp. 19-33.

Leech, Geoffrey; Mc Enery, Tony & Martin, Wynne (1997) Further levels of annotation. In Garside, Roger et al. (eds.) (1997), pp. 85-101.

Leech, Geoffrey; Myers, Greg & Jenny, Thomas (eds.) (1995) *Spoken English on Computer: Transcription, Mark-up and Application*. New York: Longman.

Lewandowska-Tomaszczyk, Barbara & Marcel, Thelen (eds.) (2002) *Translation and Meaning 6. Proceedings of the Lodz session of the 3rd international Maastricht- Lodz duo colloquium on "Translation and meaning", held in Lodz, Poland, 22-24 September 2000.* Maastricht: Hogeschool Zuyd, Maastricht School of Translation and Interpreting.

Lewandowska-Tomaszczyk, Barbara & Patrick James, Melia (eds.) (1997) *International Conference on Practical Applications in Language Corpora. Lodz, Poland, 10-14 April 1997: Proceedings*. Lodz: Lodz University Press.

Lindquist, P. Peter & Giambruno, Miguélez (2006) The MRC approach: Corpus-based techniques applied to interpreter performance analysis and instruction. *FORUM International Journal of Interpretation and Translation* 4/1, pp. 103-138.

Lindquist, P. Peter (2005) New technologies, Discourse Analysis, and the spoken word: MRC, an empirical approach to interpreter performance evaluation and pedagogy. *META* 50/1. Online: <http://id.erudit.org/iderudit/019848ar>.

Lino, Maria Teresa; Xavier, María Francisca; Ferreira, Fátima; Costa, Rute & Silva, Raquel, with the collaboration of Carla Pereira, Filipa Carvalho, Milene Lopes, Mónica Catarino, Sérgio Barros (eds.) (2004) *Proceedings of the Fourth International Conference on Language Resources and Evaluation.* Lisbon: ELRA.

Lyding, Verena (ed.) (2009) *LULCL II 2008. Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II). "Combining efforts to foster computational support of minority languages". Bozen-Bolzano, 13th-14th November 2008*. Bolzano: Accademia Europea Bolzano. Online: <http://www.eurac.edu/Org/LanguageLaw/Multilingualism/Projects/LULCL_II_proceedings.htm>.

Mack, Gabriele (2006) Detto scritto: un fenomeno, tanti nomi. *inTRAlinea* Special Issue: Respeaking. Online: <http://www.intralinea.it/specials/respeaking/eng_open1.php?id=P464>.

Mackintosh, J. (1983) *Relay Interpretation: An Exploratory Study*. Tesi di laurea non pubblicata. Birkbeck College, University of London.

MacWhinney, Brian (1997) *Il progetto CHILDES: strumenti per l'analisi del linguaggio parlato*. Edizione italiana a cura di Elena Pizzuto e Umberta Bortolini. Tirrenia, Pisa: Edizioni del Cerro.

MacWhinney, Brian (2000) *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates. Online: <http://childes.psy.cmu.edu/manuals/chat.pdf>.

Maia, Belinda (1997) Do-it-yourself corpora… with a little bit of help from your friends! In Lewandowska-Tomaszczyk, Barbara & Patrick James, Melia (eds.), pp. 403-410.

Mair, Christian & Marianne, Hundt (eds.) (2000) *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20). Freiburg im Breisgau 1999*. Amsterdam/Atlanta: Rodopi.

Malmkjaer, Kristen (ed.) (1998) *Translation and Language Teaching: Language Teaching and Translation*. Manchester: St. Jerome.

Marcos Marín, Francisco (2005) Lo que aprendimos al elaborar el corpus oral peninsular. In Burr, Elisabeth (a cura di), pp. 77-96.

Martin, Philippe (2004) WinPitch Corpus. A text to speech alignment tool for multimodal corpora. In Lino, M. et al. (eds.), pp. 537-540.

Marzocchi C. & G. Zucchetto (1997) Some considerations on interpreting in an institutional context: the case of the European Parliament. *Terminologie et Traduction* 3, pp. 70-85.

Marzocchi C. (2007) Translation-Transcription-Interpretation. Notes on the European Parliament verbatim report of proceedings. *Across Languages and Cultures* 8/2, pp. 249-254.

Matsubara, Shigeki; Takagi, Akira; Kawaguchi, Nobuo & Yasuyoshi, Inagaki (2002) Bilingual spoken monologue corpus for simultaneous machine interpretation research. In González Rodríguez, Manuel & Carmen Paz, Suárez Araujo (eds.), Volume I, pp. 153-159.

Mauranen, A. (2003) The Corpus of English as Lingua Franca in academic settings. *TESOL Quarterly* 37/3, pp. 513-527.

McEnery, Tony & Andrew, Wilson (1997) Multimedia corpora. In Lewandowska-Tomaszczyk, Barbara & Patrick James, Melia (eds.), pp. 24-33.

McEnery, Tony & Andrew, Wilson (2001) *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, Tony & Paul, Rayson (1997) A Corpus/ Annotation toolbox. In Garside, Roger et al. (eds.), pp. 194-208.

McEnery, Tony; Xiao, Richard & Yukio, Tono (2006) *Corpus-based Language Studies. An advanced resource book*. London/New York: Routledge.

Mead, Peter (2000) Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter* 10, pp. 89-101. Online: <http://www.openstarts.units.it/dspace/bitstream/10077/2451/1/05.pdf>.

Mead, Peter (2002a) Exploring hesitation in consecutive interpreting: an empirical study. In Garzone, Giuliana & Maurizio, Viezzi (eds.), pp. 73-82.

Mead, Peter (2002b) La percezione del congiuntivo italiano nell'interpretazione verso l'inglese. In Schena, L. et al. (a cura di), pp. 327-333.

Merlini, Raffaela (2007) Teaching dialogue interpreting in higher education: a research-driven, professionally oriented curriculum design. In Musacchio, Maria Tesresa & Geneviève, Henrot Sostero (eds.), pp. 278-306.

Messina, Alessandro (2001) Lingue e interpretazione. Riflessioni sull'insegnamento / apprendimento linguistico nella formazione degli interpreti di conferenza. *inTRAlinea* 4. Online: <http://www.intralinea.it/volumes/eng_open.php?id=P132>.

Meyer, Bernd & Thomas, Schmidt (s.d.) CoSi – A corpus of consecutive and simultaneous interpreting. Online: <http://www1.uni-hamburg.de/exmaralda/files/k6-korpus/CoSi.pdf>.

Meyer, Bernd (1998) What transcriptions of authentic discourse can reveal about interpreting. *Interpreting* 3/1, pp. 65-83.

Meyer, Bernd (2000) The computer-based transcriptions of simultaneous interpreting. In Dimitrova, E. B. & K. Hyltenstam (eds.), pp. 151-158.

Meyer, Bernd (2008) Interpreting proper names: different interventions in simultaneous and consecutive interpreting. *trans-kom* 1/1, pp. 105-122. Online: <http://www.trans-kom.eu/ihv_01_01_2008.html>.

Meyer, F. Charles & Gerald, Nelson (2006) Data collection. In Aarts, Bas & April, McMahon (eds.), pp. 93-113.

Mikhailov, Mikhail (2001) Two approaches to automated text aligning of parallel fiction texts. *Across Languages and Cultures* 2/1, pp. 87-96.

Milroy, Lesley & Matthew, Gordon (2003) *Sociolinguistics. Methods and Interpretation*. Malden/Oxford: Blackwell Publishing.

Monacelli, Claudia (2009) *Self-Preservation in Simultaneous Interpreting. Surviving the role*. Amsterdam/Philadelphia: John Benjamins.

Moneglia, Massimo & Emanuela, Cresti (1997) Intonazione e criteri di trascrizione del parlato. In Bortolini, Umberta & Elena, Pizzuto (a cura di), pp. 57-90.

Moneglia, Massimo & Emanuela, Cresti (2001) The value of prosody in the transition to complex utterances: Data and theoretical implications from the acquisition of Italian. In Almgren, Margareta et al. (eds.), vol. 2, pp. 850-872.

Moneglia, Massimo (2005) I corpora dell'italiano parlato di LABLITA: criteri di costituzione, unità di analisi e comparabilità dei dati linguistici orali. In Burr, E. (a cura di), pp. 213-231.

Monti, Cristina; Bendazzoli, Claudio; Sandrelli, Annalisa & Mariachiara, Russo (2005) Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus. *Meta* 50/4. Online: <http://www.erudit.org/revue/meta/2005/v50/n4/019850ar.pdf>.

Moreno Sandoval, A. & Goñi, J. M. (1995) A Morphological model and processor for Spanish implemented in Prolog. In Sessa, M.I. & M. Alpuente Frasnedo (eds.), pp. 321-331.

Moreno Sandoval, Antonio & Guirao, José María (2003) Tagging a spontaneous speech corpus of Spanish. In *Proceedings of the VI Conference on Recent Advances in Natural*

Language Processing. Online: <http://www.lllf.uam.es/ESP/Publicaciones/publicaciones2003.html>.

Moreno Sandoval, Antonio & Guirao, José María (2006). Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Kawaguchi, Yuji et al. (eds.), pp. 199-218.

Moreno Sandoval, Antonio; De la Madrid, Gillermo; Alcántara, Manuel; Gonzalez, Ana; Guirao M. José & Raúl, De la Torre (2005) The Spanish corpus. In Cresti, Emanuela & Massimo, Moneglia (eds.), pp. 135-161.

Munday, Jeremy (2008) *Style and Idelogy in Translation: Latin American Writing in English.* New York/London: Routledge.

Musacchio, Maria Tesresa & Geneviève, Henrot Sostero (eds.) (2007) *Tradurre: formazione e professione*. Padova: CLEUP

Namy, Claude (1978) Reflections on the training of simultaneous interpreters: a meta-linguistic approach. In Gerver, D. & H.W. Sinaiko (eds.), pp. 25-33.

Nencioni, G. (1989) *Saggi di lingua antica e moderna*. Torino: Rosemberg & Sellier.

Nesi, H. & P. Thompson, (2006) *The British Academic Spoken English Corpus Manual*. Online: <http://www.coventry.ac.uk/researchnet/external/content/1/c4/26/35/v1211446010/user/base_manual.pdf>.

Neuman, William Lawrence (1997) *Social Research Methods: Qualitative and Quantitative Approaches* (3rd edition). Boston: Ally & Bacon.

Niemants, Natacha Sarah Alexandra (2009) *La formazione degli interpreti di comunità. Un confronto tra interpretazioni "didattiche" e "reali"*. Descrizione del progetto di ricerca per la tesi di dottorato presso l'Università degli Studi di Modena e Reggio Emilia. Online: <http://www.dailynterpreter.com/wp-content/2009/07/phdprojectoncommunityinterpreting-it.pdf>.

O'Connell, C. Daniel & Sabine, Kowal (1994) Some current transcription systems for spoken discourse: a critical analysis. *Pragmatics* 4/1, pp. 81-107.

O'Connell, C. Daniel & Sabine, Kowal (1999) Transcription and the issue of standardization. *Journal of Psycholinguistic Research* 28/2, pp. 103-120.

O'Connell, C. Daniel & Sabine, Kowal (2009) Transcription systems for spoken discourse. In D'hondt, S. et al. (eds.), pp. 240-254.

Ochs, Elinor (1999) Transcription as theory. In Jarowski, A. & N., Coupland (eds.), pp. 167-182.

Olohan, Maeve (2004) *Introducing Corpora in Translation Studies*. London/New York: Routledge.

Ono, Takahiro; Tohyama, Hitomi & Shigeki, Matsubara (2008) Construction and analysis of word-level time-aligned simultaneous interpretation corpus. In Calzolari, N. et al. (eds.). Online: <http://www.lrec-conf.org/proceedings/lrec2008/>.

Oostdijk, N.; Goedertier, W.; Van Eynde, F.; Boves, L.; Martens, J.P.; Moortgat, M. & H. Baayen (2002) Experiences from the Spoken Dutch Corpus Project. In González Rodríguez, Manuel & Carmen Paz Suárez Araujo (eds.), pp. 340-347.

Oostdijk, N.; Kristoffersen, G. & G. Sampson (eds.) (2004) *Compiling and Processing Spoken Language Corpora, LREC 2004 Satellite Workshop, Fourth International Conference on Language Resources and Evaluation, 24th May 2004*. Lisbon: ELRA.

Orletti, Franca & Renata, Testa (1991) La trascrizione di un corpus interlingua: aspetti teorici e metodologici. *SILTA* 20/2, pp. 243-283.

Palazzi, Maria Cristina (1999) Aspetti pratici della professione. In Falbo, C. et al. (a cura di), pp. 41-59.

Parks, Gerald (a cura di) (1995) *Miscellanea n. 2*. Trieste: Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Parlamento europeo (2009) *Regolamento*. Online: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+RULES-EP+20091201+0+DOC+PDF+V0//IT&language=IT>.

Pöchhacker, Franz & Waltraud Kolb (2007) *Interpreting for the record: A case study of asylum review hearings*. Comunicazione presentata in occasione del convegno Critical Link 5 – Quality in Interpreting: A Shared Responsibility, 11-15 April 2007 Parramatta – Sydney (Australia).

Pöchhacker, Franz (1992) The role of theory in simultaneous interpreting. In Dollerup, C. & A. Lindegaard (eds.), pp. 211-220.

Pöchhacker, Franz (1994a) Quality assurance in simultaneous interpreting. In Dollerup, cay & Annette, Lindegaard (eds.) (1994), pp. 232-242.

Pöchhacker, Franz (1994b) *Simultandolmetschen als komplexes Handeln.* Tübingen: Gunter Narr.

Pöchhacker, Franz (2004) *Introducing Interpreting Studies*. London/New York: Routledge.

Pöchhacker, Franz (2006) "Going social?" On pathways and paradigms in interpreting studies. In Pym, A. et al. (eds.), pp. 215-232.

Pöchhacker, Franz (2007) *Interpreting in the Refugee and Migration Review Tribunals*. Comunicazione presentata in occasione del convegno Critical Link 5 – Quality in Interpreting: A Shared Responsibility, 11-15 April 2007 Parramatta – Sydney (Australia).

Pöchhacker, Franz (2008) The turns of Interpreting Studies. In Hansen, G. et al. (eds.), pp. 25-46.

Pöchhacker, Franz (2009) Inside the 'black box'. Can interpreting studies help the profession if access to real-life settings is denied? *The Linguist* 48/2, pp. 22-23.

Psathas, George & Timothy, Anderson (1990) The 'practices' of transcription in conversation analysis. *Semiotica* 78/1-2, pp. 75-99.

Pusch, D. Claus & Wolfgang, Raible (eds.) (2002) *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache. – Romance Corpus Linguistics: Corpora and Spoken Language*. Tübingen: Narr.

Pusch, D. Claus (2002) A survey of spoken language corpora in Romance. In Pusch, D. Claus & Wolfgang, Raible (eds.), pp. 245-264.

Pym, Anthony; Shlesinger, Miriam & Zuzana, Jettmarova (eds.) (2006) *Sociocultural Aspects of Translating and Interpreting.* Amsterdam/Philadelphia: John Benjamins.

Quasthoff, Uta M. (ed.) (1995) *Aspects of Oral Communication*. Berlin/New York: Walter de Gruyter.

Räisänen, Christine (1999) *The conference Forum as System of Genres. A Sociocultural Study of Academic Conference Practices in Automotive Crash-Safety Engineering*. Göteborg: Acta Universitatis Gothoburgensis.

Recalde, Monserrat & Victoria, Vázquez Rozas (2009) Problemas metodológicos en la formación de corpus orales. In Cantos Gómez, P. & A. Sánchez Pérez (eds.), pp. 37-49.

Riccardi, Alessandra (1995) La conferenza quale evento comunicativo ed il ruolo dell'interprete. In Parks, G. (a cura di), pp.99-104.

Riccardi, Alessandra (1997) Lingua di conferenza. In Gran, L. & A. Riccardi (a cura di), pp. 59-74.

Riccardi, Alessandra (2003) *Dalla traduzione all'interpretazione. Studi d'interpretazione simultanea*. Milano: LED.

Riccardi, Alessandra (2009) La ricerca in interpretazione simultanea – Verifica sperimentale di dati empirici. In Cavagnoli, Stefania et al. (eds.), pp. 359-373.

Rowley-Jolivet, Elizabeth & Shirley, Carter-Thomas (2005) The rhetoric of conference presentation introductions: context, argument and interaction. *International Journal of Applied Linguistics* 15/1, pp. 45-70.

Russell, Debra & Sandra, Hale (eds.) (2008) *Interpreting in Legal Settings*. Washington: Gallaudet University Press.

Russo, Mariachiara (1990) Disimetrías y actualización: un experimento de interpretación simultanea (español-italiano). In Gran, L. & C., Taylor (eds), pp.158-225.

Russo, Mariachiara (1997) Morphosyntactical asymmetries between Spanish and Italian and their effect during simultaneous interpretation. In Klaudy, K. & J. Kohn (eds.), pp. 268-272.

Russo, Mariachiara (1999) La conferenza come evento comunicativo. In Falbo, C. et al. (a cura di), pp. 87-102.

Russo, Mariachiara (2007) European Parliament Interpreting Corpus (EPIC): rasgos distintivos de la interpretación simultánea de los discursos en español. *Rivista di Filologia e Letterature Ispaniche* 10, pp. 289-304.

Russo, Mariachiara (2008) *Information processing patterns in simultaneous interpreting from Spanish into Italian: A corpus-based study*. Paper presented at the 41th Annual Meeting of the Societas Linguistica Europea "Languages in Contrast", Forlì 17-20 September.

Russo, Mariachiara (2010) Reflecting on interpreting practice: graduation theses based on the European Parliament Interpreting Corpus (EPIC). In Zybatow N. L. (ed.), pp. 35-50.

Russo, Mariachiara; Bendazzoli, Claudio; Sandrelli, Annalisa & Nicoletta, Spinolo (2010) *The European Parliament Interpreting Corpus (EPIC): Implementations and Developments*. Comunicazione presentata al convegno *Emerging Topics in Translation and Interpreting - Nuovi percorsi in traduzione e interpretazione*, Università di Trieste, 16-18 giugno 2010.

Russo, Mariachiara; Claudio Bendazzoli & Annalisa Sandrelli (2006) Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC (European Parliament Interpreting Corpus). *Forum* 4/1, pp. 221-254.

Sandrelli, Annalisa & Claudio Bendazzoli (2005) Lexical patterns in simultaneous interpreting: a preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series* 1/ 1, ISSN 1747-9398. Online: <www.corpus.bham.ac.uk/PCLC>.

Sandrelli, Annalisa & Claudio, Bendazzoli (2006) Tagging a corpus of interpreted speeches: The European Parliament Interpreting Corpus (EPIC). In *Proceedings of the LREC 2006 Conference, Genova, Magazzini del Cotone 24-26 May 2006*. Genova: ELRA. Online: <http://hnk.ffzg.hr/bibl/lrec2006/>.

Sandrelli, Annalisa & Jesús, de Manuel Jerez (2007) The impact of Information and Communication Technology on interpreter training. *The Interpreter and Translator Trainer* 1/2, pp. 269-303.

Sandrelli, Annalisa (2002) Computers in the training of interpreters: Curriculum design issues. In Garzone, G. et al. (eds.), pp. 189-204.

Sandrelli, Annalisa (2003a) Herramientas informáticas para la formación de intérpretes: Interpretations y The Black Box. In de Manuel Jerez, Jesús (coord.) (2003), pp. 67-112.

Sandrelli, Annalisa (2003b) New Technologies in Interpreter Training: CAIT. In Gerzymisch-Arbogast, Heidrun et al. (eds.) (2003), pp. 261-293.

Sandrelli, Annalisa (2010) Corpus-based Interpreting Studies and interpreter training: A modest proposal. In Zybatow, N. L. (ed.).

Sandrelli, Annalisa; Bendazzoli, Claudio & Mariachiara, Russo (2007) *The impact of topic, mode and speed of delivery on the interpreter's performance: a corpus-based quality evaluation*. Poster presentato al convegno Critical Link 5. *Quality in interpreting: a shared responsibility*, Parramatta – Sydney (Australia).

Sandrelli, Annalisa; Bendazzoli, Claudio & Mariachiara, Russo (2010) European Parliament Interpreting Corpus (EPIC): Methodological issues and preliminary results on lexical patterns in simultaneous interpreting. *International Journal of Translation* 22/1-2, pp. 165-203.

Sankoff, Gillian (1980a) A quantitative paradigm for the study of communicative competence. In Sankoff Gillian (ed.) (1980), pp. 47-79.

Sankoff, Gillian (ed.) (1980) The *Social Life of Language*. Philadelphia: University of Pennsylvania Press.

Schena, Leo; Prandi, Michele & Marco, Mazzoleni (a cura di) (2002) *Intorno al congiuntivo*. Bologna: Clueb.

Schiffrin, Deborah; Tannen, Deborah & E. Heidi, Hamilton (eds.) (2001) *The Handbook of Discourse Analysis*. Malden, MA: Blackwell.

Schjoldager, Anne (1995) An Exploratory Study of Translational Norms in Simultaneous Interpreting: Methodological Reflections. *Hermes, Journal of Linguistics* 14, pp. 65-87. Online: <http://download2.hermes.asb.dk/archive/download/H14_05.pdf>.

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing. September 1994*. Online: <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>.

Schmid, Helmut (1995) Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop. March 1995*. Online: <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf>.

Schmidt, Thomas (2001) The transcription system EXMARaLDA: An application of the annotation graph formalism as the basis of a database of multilingual spoken discourse. In: Bird, S. et al. (eds.), pp. 219-227. Online: <http://www.ldc.upenn.edu/annotation/database/papers/Schmidt/2.2.schmidt.pdf>.

Schmidt, Thomas (2003) Visualising linguistic annotation as interlinear text. Arbeiten zur Mehrsprachigkeit, - Working Papers in Multilingualism B/46. Online: <http://www1.uni-hamburg.de/exmaralda/files/Visualising-final.pdf>.

Schmidt, Thomas (2004) Transcribing and annotating spoken language with EXMARaLDA. *Proceedings of the LREC-Workshop on XML-based richly annotated corpora, Lisbon 2004*. Paris: ELRA. Online: <http://www1.uni-hamburg.de/exmaralda/files/Paper_LREC.pdf>.

Schmidt, Thomas (2009) Creating and working with spoken language corpora in EXMARaLDA. In Lyding, Verena (ed.), pp. 151-164.

Schneider, Stefan (2002) An online databse version of the LIP corpus. In Pusch, D. Claus & Wolfgang, Raible (eds.), pp. 201-208.

Schweda Nicholson, N. (1990) The role of shadowing in interpreter training. *The Interpreters' Newsletter* 3, pp. 33-40.

Scott, Mike (2003) *Wordsmith Tools, Computer Software*. Oxford: Oxford University Press.

Seleskovitch, Danica (1978) Language and cognition. In Gerver, D. & H. V. Sinaiko eds, pp. 333-341.

Sessa, M.I. & M. Alpuente Frasnedo (eds.) (1995) *Gulp-Prode '95: Joint Conference on Declarative Programming, Marina di Vietri sul Mare, Italy, 11-14 september, 1995*. Salerno: Poligraf Press.

Setton, Robin (1999) *Simultaneous Interpretation: A Cognitive-Pragmatic Analysis*. Amsterdam/Philadelphia: John Benjamins.

Setton, Robin (2002) A methodology for the analysis of interpretation corpora. In Garzone, Giuliana & Maurizio, Viezzi (eds.), pp. 29-45.

Setton, Robin (s.d.) *Corpus-based interpretation studies (CIS): reflections and prospects*. Paper delivered at the Symposium on Corpus-based Translation Studies: Research and Applications, Pretoria, July 22-25, 2003. La versione aggiornata di questo articolo è in corso di stampa in Kruger, Alet; Walmach, Kim & Jeremy, Munday (eds.) (2011) *Corpus-based Translation Studies Research and Applications.* London/New York: Continuum.

Shalom, N. Celia (1995) The discourse management role of the Chair in academic conference presentation sessions. *Interface: Journal of Applied Linguistics* 10/1, pp. 47-62.

Shalom, N. Celia (2002) The academic conference: a forum for enacting genre knowledge. In Ventola et al. (eds.), pp. 51-68.

Shlesinger, Miriam & Noam, Ordan (2010) *Interpreting as a genre unto itself. Findings based on a tagged Hebrew corpus*. Comunicazione presentata in occasione della conferenza *The use of corpora in Interpreting Studies*, 17/02/2010 Forlì, SSLMIT.

Shlesinger, Miriam (1994) Intonation in the production of and perception of simultaneous interpretation. In Lambert, Sylvie & Barbara Moser-Mercer (eds.), pp. 225-236.

Shlesinger, Miriam (1998) Corpus-based Interpreting Studies as an offshoot of Corpus-based Translation Studies. *Meta* 43/4, pp. 486-493.

Shlesinger, Miriam (2003) Effects of presentation rate on working memory in simultaneous interpreting. *The Interpreters' Newsletter* 12, pp. 37-49.

Shlesinger, Miriam (2008) Towards a definition of Interpretese. An intermodal, corpus-based study. In G. Hansen, Chesterman, A. & H. Gerzymisch-Arbogast (eds.), pp. 237-253.

Silverman, David (ed.) (1997) *Qualitative Research: Theory, Method and Practice*. London: SAGE Publications.

Simpson, C. Rita; Lee, Y.W. David & Leicher, Sheryl. Revised by Annelie Ädel (2007) *MICASE Manual. The Michigan Corpus of Academic Spoken English. Version 3, Work in Progress, dated June 8, 2007*. Ann Arbor, Michigan, USA: English Language Institute, The University of Michigan. Online: < http://micase.elicorpora.info/micase-manual-pdf>.

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

Sinclair, John (1995) From theory to practice. In Leech, Geoffrey et al. (eds.), pp. 99-109.

Sinclair, John (ed.) (2004) *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Snell-Hornby, Mary; Pöchhacker, Franz & Klaus, Kaindl (eds.) (1994) *Translation Studies: An Interdiscipline*. Amsterdam/Philadelphia: John Benjamins.

Somers, H. (ed.) (1996) *Terminology, LSP and Translation.* Amsterdam/Philadelphia: John Benjamins.

Sornicola, Rosanna (1981) *Sul parlato*. Bologna: Il Mulino.

Sornicola, Rosanna (1984) Sulla costituzione dei testi parlati. In Coveri, Lorenzo (a cura di), pp. 341-350.

Spina, Stefania (2005) Il Corpus di Italiano Televisivo (CIT): struttura e annotazione. In Burr, E. (ed.), pp. 413-426.

St.John, E. & M. Chattle, (1998) Multiconcord: The Lingua Multilingual Parallel Concordancer for Windows. *ReCALL Newsletter* 13, pp. 7-9.

Stame, Stefania (ed.) (1997) *Psycholinguistics as a Multidisciplinarily Connected Science*. Vol. 2. Cesena: Ponte Vecchio.

Straniero Sergio, Francesco (2007) *Talkshow Interpreting. La mediazione linguistica nella conversazione spettacolo*. Trieste: Edizioni Universitare Trieste.

Stubbs, M. (1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford and Cambridge, MA: Blackwell.

Sunnari, M. (1997) Finnish interpreting services in the European Union after the first year. In Klaudy, K. & J. Kohn (eds.), pp. 87-90.

Sunnari, Marianna (1999) Return interpreting. A dual responsibility. In Álvarez Lugrís, A. & A. Fernández Ocampo (coord.), pp. 317-320.

Svartvik, Jan (ed.) (1990) *The London-Lund corpus of spoken English: Description and Research*. Lund: Lund University Press.

Szakos, Josef & Ulrike Glavitsch (2004a) Portability, modularity and seamless speech-corpus indexing and retrieval: a new software for documenting (not only) the endangered Formosan Aboriginal languages. *Proceedings of the E-MELD Language Digitization Project Conference, Workshop on Linguistic Databases and Best Practice, Detroit, USA, July 15 - 18, 2004*. Online: <http://emeld.org/workshop/2004/szakos-paper.html>.

Szakos, Josef & Ulrike Glavitsch (2004b) *Seamless speech indexing and retrieval: developing an new technology for the documentation and retrieving of endangered Formosan languages*. Comunicazione presentata al convegno Sixth International Conference on Teaching and Language Corpora (TALC6). Università di Granada (Spagna) 4-7 luglio 2004.

Takagi, Akira; Matsubara, Shigeki; Matsubara, Shigeki & Yasuyoshi, Inagaki (2002) A corpus-based analysis of simultaneous interpretation. *Proceedings of International Joint Conference of the 5th Symposium on Natural Language Processing* (SNLP-2002), pp.167-174. Online: <http://slp.itc.nagoya-u.ac.jp/web/papers/2002/snlp2002_takagi.pdf>.

Taylor Torsello, C.; Brunetti, Giuseppe & Nicoletta, Penello (eds.) (2001) *Corpora testuali per ricerca, traduzione e apprendimento linguistico*. Padova: Unipress.

Thompson, Paul (2005) Spoken language corpora. In Wynne, M. (ed.). Online: <http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm/>.

Timarová, Šárka (2005) Corpus Linguistics methods in Interpreting research: A case study. *The Interpreters's Newsletter* 13, pp. 65-70. Online: <http://www.openstarts.units.it/dspace/bitstream/10077/2471/1/05.pdf>.

Tirkkonen-Condit, Sonis & Riitta, Jääskeläinen (eds.) (2000) *Tapping and Mapping the Process of Translation and Interpreting*. Amsterdam/Philadelphia: John Benjamins.

Tohyama, Hitomi & Shigeki, Matsubara (2006) Development of web-based teaching material for simultaneous interpreting learners using bilingual speech corpus. In *Proceedings of ED-MEDIA 2006: World Conference on Educational Multimedia and Hypermedia, Orlando, Florida, USA, June 26-30*, pp. 2906-2911. Online: <http://slp.itc.nagoya-u.ac.jp/web/papers/2006/EDMEDIA06_tohyama.pdf >.

Tohyama, Hitomi; Matsubara, Shigeki; Kawaguchi, Nobuo & Yasuyoshi, Inagaki (2005) Construction and utilization of bilingual speech corpus for simultaneous machine interpretation research. In P*roceedings of 9th European Conference on Speech Communication and Technology (Eurospeech-2005)*, pp. 1585-1588. Online: <http://slp.itc.nagoya-u.ac.jp/web/papers/2005/eurospeech2005_tohyama_final.pdf >.

Torresi, Ira (2009) Sociolinguistics in Interpreting research. In Cavagnoli, Stefania et al. (eds.), pp. 390-404.

Ulrych, Margherita (1997) The impact of multilingual parallel concordancing on translation. In Lewandowska-Tomaszczyk, Barbara & Patrick James, Melia (eds.), pp. 421-435.

Ulrych, Margherita (2001) What corpora for what purpose: a translation-based view. In Taylor Torsello, C. et al. (eds.), pp. 361-372.

Valentini, Cristina (2009) *Creazione e sviluppo di corpora multimediali. Nuove metodologie di ricerca nella traduzione audiovisiva*. Tesi di dottorato, Università di Bologna, sede di Forlì.

Ventola, Eija; Shalom, Celia & Susan Thompson (eds.) (2002) *The Language of Conferencing*. Frankfurt am Main: Peter Lang GmbH.

Viezzi, Maurizio (1999) Interpretazione simultanea: attività specifica per coppie di lingue? *Settentrione* 11/1, pp. 133-159.

Viezzi, Maurizio (2002) La non-selezione del congiuntivo quale opzione strategica nell'interpretazione simultanea dall'inglese in italiano. In Schena, L. et al. (a cura di), pp. 350-359.

Vik-Tuovinen, Gun-Viol (2000) The interpreters' comments in interpreting situations. In Tirkkonen-Condit, S. & R. Jääskeläinen (eds.), pp. 18-26.

Vuorikoski Anna-Riitta (2004) *A Voice of its Citizens or a Modern Tower of Babel? The Quality of Interpreting as a Function of Political Rhetoric in the European Parliament*. Tampere: Tampere University Press. Online: <http://acta.uta.fi/teos.php?id=9744>.

Wadensjö, Cecilia (1998) *Interpreting as interaction*. London/New York: Longman.

Wallmach, Kim (2002a) "Seizing the surge of language by its soft, bare skul": simultaneous interpreting, the Truth Commission and *Country of My Skull*. *Current Writing* 14/2, pp. 64-82.

Wallmach, Kim (2002b) Using parallel corpora to determine interpreting strategies for languages of limited diffusion in South Africa. In Lewandowska-Tomaszczyk, Barbara & Marcel, Thelen (eds.), pp. 503-509.

Wallmach, Kim (2004) 'Pressure players' or 'choke artists'? How do Zulu simultaneous interpreters handle the pressure of interpreting in a legislative context? *Language Matters* 34, pp. 179-200. Online: <http://www.multilingua.co.za/pdfs/Wallmach_2004_Language_Matters.pdf>.

Webber, Pauline (1999) Public discourse in science: a comparison between English and Italian lectures. In Azzaro, G. & M. Ulrych (a cura di), pp. 113-126.

Webber, Pauline (2004) The use of a spoken corpus for the analysis of academic conference monologues. In Bondi et al. (eds.), pp. 87-104.

Wilson, Andrew & Jenny, Thomas (1997) Semantic annotation. In Garside, Roger et al. (eds.), pp. 53-65.

Wynne, M. (2005a) Archiving, distribution and preservation. In Wynne, M. (ed.) (2005). Online: <http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter6.htm>.

Wynne, M. (ed.) (2005) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. Online: <http://ahds.ac.uk/linguistic-corpora/>.

Yagi, M. Sane (1994) *A Psycholinguistic Model for Simultaneous Translation, and Proficiency Assessment by automated Acoustic Analysis of Doscourse*. Tesi di dottorato, The University of Auckland (New Zealand).

Yagi, M. Sane (1999) Computational discourse analysis for interpretation. *Meta* 44/2, pp. 268-279. Online: <http://www.erudit.org/revue/meta/1999/v44/n2/004627ar.pdf>.

Yuste, Elia Rodrigo (ed.) (2008) *Topics in Language Resources for Translation and Localisation*. Amsterdam/Philadelphia: John Benjamins.

Zanettin, Federico (1998) Bilingual comparable corpora and the training of translators. *Meta* 43/4, pp. 616-630. Online: <http://id.erudit.org/iderudit/004638ar>.

Zanettin, Federico (2001) IperGrimus. Ipermedia e traduzione. *inTRAlinea* Ipermedia. Online: <http://www.intralinea.it/intra/ipermedia/IperGrimus/_private/default.htm>.

Zanettin, Federico; Bernardini, Silvia & Dominic, Stewart (eds.) (2003) *Corpora in Translator Education*. Manchester/Northampton: St Jerome.

Zannirato, Alessandro (2008) Teaching interpreting and interpreting teaching: a conference interpreter's overview of second language acquisition. In Di Kearns, John (ed.), pp. 19-38.

Zorzi, Daniela (2004) Studi conversazionali e interpretazione. In Bersani Berselli, G. et al. (eds.), pp. 73-89.

Zybatow, N. Lev (ed.) (2010) *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungen zur Translationswissenschaft VI (Forum Translationswissenschaft, Band 12)*. Frankfurt am Main [etc.]: Peter Lang.

Sitografia

Tutti i collegamenti ipertestuali sono aggiornati al 1 dicembre 2010.

AIIC (Association Internationale des Interprètes de Conférence)
http://www.aiic.net/

BADIP (Banca Dati dell'Italiano Parlato)
http://badip.uni-graz.at/

Baroni Marco (pagina web personale)
http://clic.cimec.unitn.it/marco/

BASE (British Academic Spoken English)
http://wwwm.coventry.ac.uk/researchnet/base/Pages/BASE.aspx

Biblioteca Multimediale del Parlamento europeo
http://www.europarl.europa.eu/eplive/archive/default_it.htm

CANCODE
http://www.cambridge.org/elt/corpus/corpora_cancode.htm

Code-A-Text C-I-SAID
(Code-A-Text Integrated System for the Analysis of Interviews and Dialogues)
http://www.code-a-text.co.uk/over_view_of_cisaid.htm

COSIH (Corpus of Spoken Israeli Hebrew)
http://www.tau.ac.il/humanities/semitic/cosih.html

CREA (Corpus de referencia del español actual)
http://corpus.rae.es/creanet.html

Czech spoken corpus (e progetti correlati)
http://ucnk.ff.cuni.cz/english/struktura.php

DAVID (Digital Audio Video Database – Università di Praga)
https://david.ff.cuni.cz/index.php

ELISA (English Language Interview Corpus as a Second-Language Application)
http://www.uni-tuebingen.de/elisa/html/elisa_index.html

ELRA (European Language Resources Association)
http://catalog.elra.info/index.php

EPIC (European Parliament Interpreting Corpus)
http://sslmitdev-online.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C.

EPIC (Interfaccia sviluppato presso la Bar-Ilan University)
http://epic.sslmit.unibo.it/

EXAKT (EXMARaLDA Analysis and Concordancing Tool)
http://www.exmaralda.org/en_exakt.html

EXMARaLDA (Extensible Markup Language for Discourse Annotation)
http://www.exmaralda.org/en_index.html

ICE International Corpus of English
http://ice-corpora.net/ice/index.htm

InterMed (Conference Interpreters Medical Sciences)
http://www.intermed.interpreters.it

Kompozer (web authoring system)
http://kompozer.net/

LDC (Linguistic Data Consortium)
http://www.ldc.upenn.edu

LLI-UAM (Laboratorio de Lingüística Informática - Universidad Autónoma de Madrid)
http://www.lllf.uam.es

Manuale interistituzionale di convenzioni redazionali dell'Unione Europea
http://publications.europa.eu/code/it/it-000100.htm

MICASE (Michigan Corpus of Academic Spoken English)
http://micase.elicorpora.info/

MIT World
http://mitworld.mit.edu/

MultiConcord
http://artsweb.bham.ac.uk/pking/multiconc/lingua.htm

ParaConc
http://www.athel.com/para.html

Parlare Italiano (osservatorio dell'italiano parlato)
http://www.parlaritaliano.it [in manutenzione]

PE (Parlamento europeo)
http://www.europarl.europa.eu/

Putty (SSH client)
http://www.chiark.greenend.org.uk/~sgtatham/putty/

Santa Barbara Corpus of Spoken American English
http://www.linguistics.ucsb.edu/research/sbcorpus.html

SIDB (Nagoya University Simultaneous Interpretation Database)
http://sidb.el.itc.nagoya-u.ac.jp/en.html

SoundWriter
http://www.linguistics.ucsb.edu/projects/transcription/tools.html

Speech Repository (DG Interpretazione – Commissione europea)
http://multilingualspeeches.tv

SpeechIndexer
http://nativesystems.inf.ethz.ch/Main/UlrikeGlavitschSoftware

TED (Technology, Entertainment, Design)
http://www.ted.com/

TEI (Text Encoding Iniziative)
http://www.tei-c.org/index.xml

UN Webcast Archives
http://www.un.org/webcast/archive.htm

Videolectures.net
http://videolectures.net/

VOICE (Vienna-Oxford International Corpus of English)
http://www.univie.ac.at/voice/

VoiceWalker
http://www.linguistics.ucsb.edu/projects/transcription/tools.html

WinPitch
http://www.winpitch.com/

WordSmith Tools
http://www.lexically.net/wordsmith/

XAIRA (XML Aware Indexing and Retrieval Architecture)
http://www.oucs.ox.ac.uk/rts/xaira/

Ringraziamenti

Si ringraziano le seguenti istituzioni per il loro prezioso sostegno:

ARCO	Cesena
Associazione Malattie Rare "Mauro Baschirotto" B.I.R.D. Europe Foundation onlus	Costozza (Vicenza)
Dipartimento di Studi Interdisciplinari su Traduzione, Lingue e Culture Università di Bologna, sede di Forlì	Forlì
Dipartimento di Politica, Istituzioni, Storia Università di Bologna	Bologna
Fondazione per la Ricerca sulla Fibrosi Cistica	Verona
ICSIM - Istituto per la Cultura e la Storia d'Impresa "Franco Momigliano"	Terni
International Association for Ambulatory Surgery Universtià di Padova	Padova
Istituto per le tecnologie della costruzione (Itc) del Cnr	Roma
Scuola Superiore di Lingue Moderne per Interpreti e Traduttori Università di Bologna, sede di Forlì	Forlì
Scuola Superiore di Studi Umanistici Università di Bologna	Bologna

Un sentito ringraziamento è dedicato a tutti i membri del gruppo di ricerca *Directionality Research Group*: Mariachiara Russo, Annalisa Sandrelli, Cristina Monti, Marco Baroni, Gabriele Mack, Elio Ballardini, Peter Mead e Silvia Bernardini, assieme a tutte le persone che hanno condiviso con generosità il loro sapere, tra cui Daniela Zorzi, Guy Aston, María Isabel Fernández García, Sara Castagnoli, Sara Piccioni, Cristina Valentini, Eros Zanchetta, Lorenzo Piccioni, Antonio Moreno Sandoval, José María Guirao, Marta Garrote Salazar, María Cristina Tovar Pérez, Leonardo Campillos e Ana Valverde Mateos. Grazie di cuore anche a Rosa Maria Bollettieri Bosinelli per il suo sostegno al progetto EPIC, a Maria Giovanna Biscu e ad Alessandra de Michele per l'impagabile assistenza nelle ultime fasi di revisione di questo volume. Infine, profonda riconoscenza e stima vanno a tutte le interpreti e a tutti gli interpreti che hanno partecipato alla ricerca: senza la vostra disponibilità, tutto questo lavoro non sarebbe stato possibile. Ve ne sono estremamente grato, così come lo saranno tutti coloro che si interesseranno al nostro mestiere.

Claudio Bendazzoli

Forlì, 1 dicembre 2010