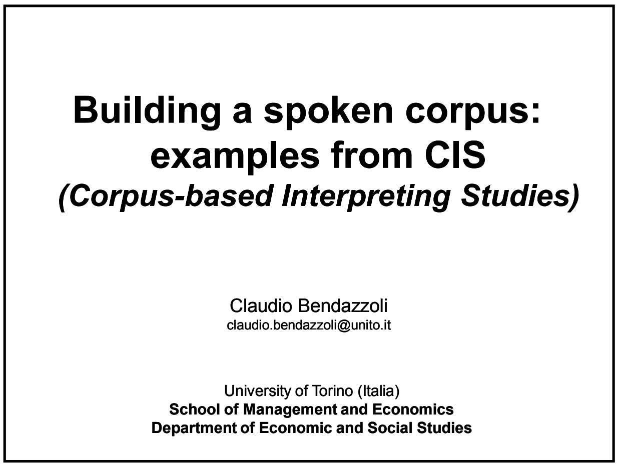
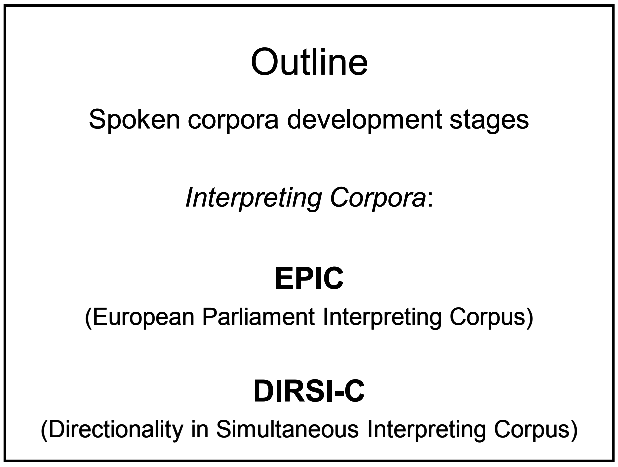


Building a spoken corpus: examples from CIS

(Corpus-based Interpreting Studies)

Claudio Bendazzoli
claudio.bendazzoli@unito.it

University of Torino (Italia)
School of Management and Economics
Department of Economic and Social Studies



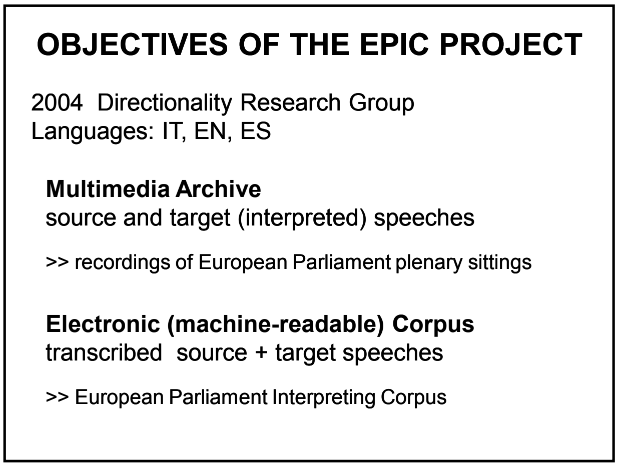
Outline

Spoken corpora development stages

Interpreting Corpora:

EPIC
(European Parliament Interpreting Corpus)

DIRSI-C
(Directionality in Simultaneous Interpreting Corpus)



OBJECTIVES OF THE EPIC PROJECT

2004 Directionality Research Group
Languages: IT, EN, ES

Multimedia Archive
source and target (interpreted) speeches

>> recordings of European Parliament plenary sittings

Electronic (machine-readable) Corpus
transcribed source + target speeches

>> European Parliament Interpreting Corpus

DIRSI-C

(Directionality in Simultaneous Interpreting)

- Italian / English
- Simultaneous interpreting (B>A + A>B)
- International conferences in Italy
  - 3 medical conferences in DIRSI-C (CFF4, ELSA, CFF5)
  - 14 conferences DIRSI-MA

Operational terminology

CL

CTS

(Baker 1995, Laviosa 1998)

CIS

(Shlesinger 1998)

ST + TT +

CTS (Laviosa 2011)

- 1993–1995
  the dawn of corpus-based translation studies
- 1996–1999
  the establishment of corpora in TS
- from 2000 onwards
  the spread of corpora across languages and cultures

…"CIS is still a cottage industry" (Setton 2011, 34)

Table 2.1 Studies of interpreting based on authentic corpora

Languages: EN English, ES Spanish, DE German, FI Finnish, FR French, HEB Hebrew, IT Italian, RU Russian, SW Swedish, TR Turkish, ZH (Standard) Chinese

Researcher	Languages	Event	Mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Oléron and Nanpon 1965	EN, FR, DE, ES	UNESCO impromptu (non-tech discussion)|	SI	pros	~ 7 minutes	Unknown	Unknown	Time lag, speed, fidelity
Dejean Le Feal 1978	FR>DE	Various speeches	SI	pros		77 pages, speakers and interpreters	Text on micro fiches, AIIC*	Recited vs. impromptu input (école du sens)
Chernov 1978	EN, FR, ES, RU	UN 1968	SI	pros	'~40 hours'	'Parallel transcripts' extracts published	Probably N. A.	Illustrate theory (redundancy & prediction)
	EN>RU, ES, FR	1978 UN satellite interpreting experiment	SI	pros		(ditto)	Probably no longer available	(ditto)
Lederer 1981	DE > FR	Railway Consortium and lab (2nd versions)	SI	2 pros	3 hours taped (original+ 2 interpreters)	63 minutes original DE, FR; some extracts synchronized (interlinear)		Illustrate theory (école du sens)
Shlesinger 1989	HEB><EN	Courtroom testimony	SI	pros	4 hours			
	FR><EN	Extracts from 2 meetings 1986–88	SI	pros	~ 4 hours	50 pp SI (2 events), fluent text speaker + interpreter, some extracts synchronized (interlinear)	No longer available	Fidelity examples (école du sens)

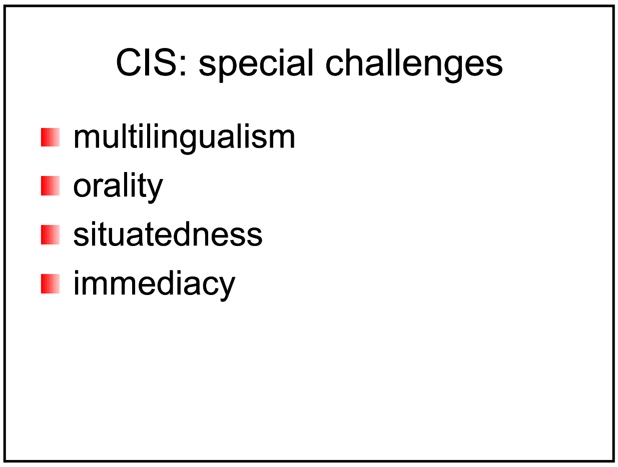
Continued

Table 2.1 Continued

Researcher	Languages	Event	Mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Pöchhacker 1994	EN><DE, FR >DE	Vienna small business conference *ICSB*	SI	pros	14 hours (original + interpreter)	Available as vol. 3 of doctoral diss. from U. of Vienna library; parts published in Pöchhacker 94	Tapes available from author	Intertextuality, situational & delivery factors (speed, slips/shifts, hesitation EN><DE)
Kalina 1998	EN><DE<>FR	*Bertelt:* 1989 public lecture (anti nuclear)	SI	6 students	70 minutes	5 versions tiered. 5s per line.	Consult, loan at Heidlbrg	Choose examples, identify errors, strategies
	DE, FR, EN	*Würzburg:* 1992 law symposium	SI	6 pros (2 per booth)		1-track audio (simulation: 2 track)	(ditto)	(ditto)
Setton 1997, 99	DE>EN	Extracts from Kalina *Würzburg* corpus	SI	1 pro + 2 in mock	14/30 minutes microanalysis	1-track audio (simulation: 2 track)	Available from author	Linguistic, cognitive-pragmatic analysis for process modelling
Wallmach 2000	EN, Zulu, Afrikaans, Sepedi	Parliament speeches Gauteng (S. Africa) Provincial legislature	SI	16 pros	Pilot: 6–8 hours EN, Afrikaans, Zulu	Pilot material transcribed (6–8 hrs)	110 hours on tape (3 languages)	Norms/strategies vs. user expectations, effect of speed, technicality on performance, language-specific strategies
Cencini 2000	EN><IT	Television Interpreting Corpus (TIC),			36,000 words	Not available for outside use		TEI Standard (computer query-able)
Fumagalli 1999–2000	EN>IT	Comparable EN(18) and IT source speeches;	CI			Not available for outside use		Features of 'interpretese' *MultiConcord-Parallel Concordancer*
Diriker 2001	EN><TR	Conference on Metaphysics and Politics (2 days)	SI	pros	150 pp transcripts	Available as Annex to Ph.D.	Bogazici Univ. Library	Interactional and sociological

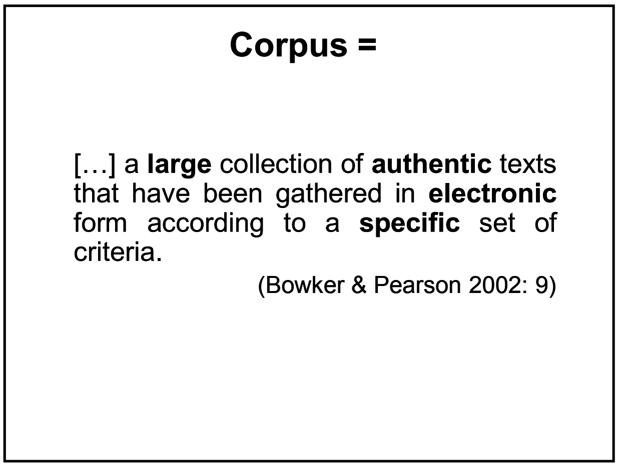
Vuorikoski 2004	mostly DE, EN ><FI, SW	European Parliament debates	SI	Ca. 70 pros	120 speeches, 65 analysed	Selected transcripts appended to PhD	CD and website	Difficulty, quality in rhetorical (political) speech genre
Beaton 2007		European Parliament debates via EBS	SI		7 hours			Text-discourse-ideology link through cohesive devices
Monacelli 2005, 2009	IT><EN	10 speeches from 4 events conferences	SI	10 pros	2 hours originals + interpretation	Selected transcripts in PhD	CD ST, TT and synchronized (from audio cues)	Interactional politeness, 'face', through deixis, mood, transitivity
Straniero 2003, 2007	IT><EN	TV talk shows	SI and CI	11 TV pros	>80 samples (1997–2002)			Emergency strategies, TV-specific quality norms
ECIS Group (Univ. of Granada)	EN>DE, ES, FR, DE, FR>ES	European Parliament debates via EbS	SI	pros	43 speeches, 73 interpretations	Linked files in MS Access, tool developed for multivariate visualization	Available from authors	Quality minimizers and maximizers, verbal and non-verbal
CIAIR (Nagoya University)	EN><JA	Monologue SI (simple lectures) and travel dialogues	SI and CI	SI: 4 pros per speech	182 hrs (1m words)	Web access/interface	Not shared	Strategies, lag, etc. for machine SI research. Own software.
EPIC**: Bologna-Forli group	EN-IT-ES (9 sub-corpora)	European Parliament sessions	SI	Multiple pros	Open-ended; 280 hrs usable in archive; ca. 280, 000 words transcribed as of mid-2010	.Available for query on line (POS-tagged and indexed database)	As separate files, on request	As of 2010: directionality, lexis, disfluencies; Ongoing (+ grad theses). Semi-automatic transcription technique

Continued



CIS: special challenges

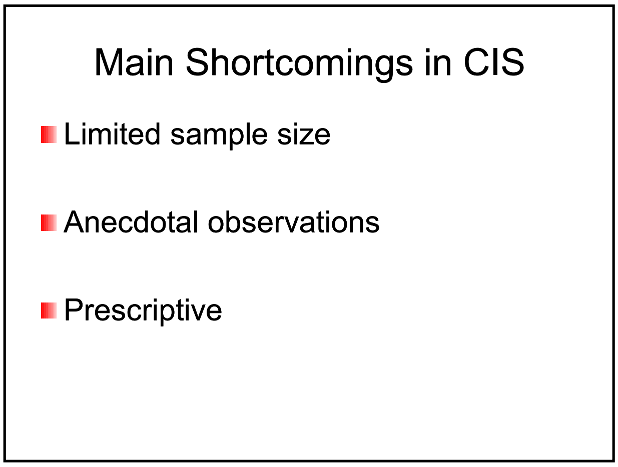
- multilingualism
- orality
- situatedness
- immediacy



Corpus =

[…] a **large** collection of **authentic** texts that have been gathered in **electronic** form according to a **specific** set of criteria.

(Bowker & Pearson 2002: 9)



Main Shortcomings in CIS

- Limited sample size
- Anecdotal observations
- Prescriptive

Developing spoken corpora:

Information Technology
Access to target populations (data)

time-consuming
challenging
enormous investment in time and painstaking manual transcription work

Developing SI corpora

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

(Thompson 2005, Bendazzoli 2010)

Methodological challenges

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

(Armstrong 1997; Baker 1996; Cencini 2002; Gile 1994, 1998, 2000; Kalina 1994; Setton 2011; Shlesinger 1998)

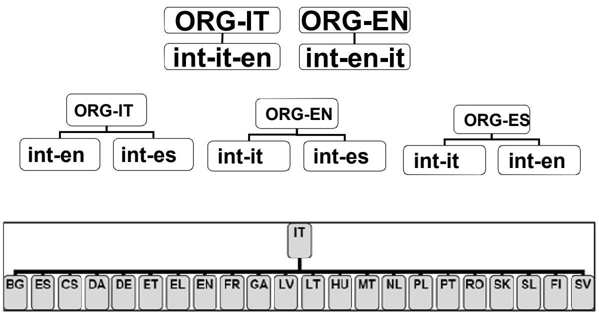
Methodological challenges

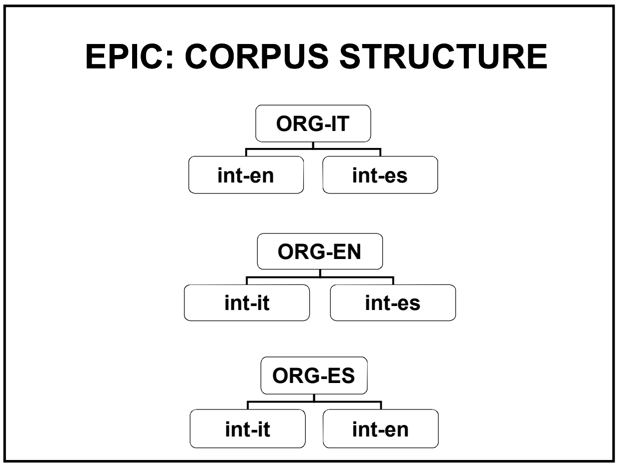
1. Corpus design >> corpus structure
   >> representativeness
2. Data collection
   >> accessibility
   >> informed consent
   >> recording

- Settings:

- Sources:

CORPUS DESIGN: structure





EPIC: CORPUS STRUCTURE

ORG-IT

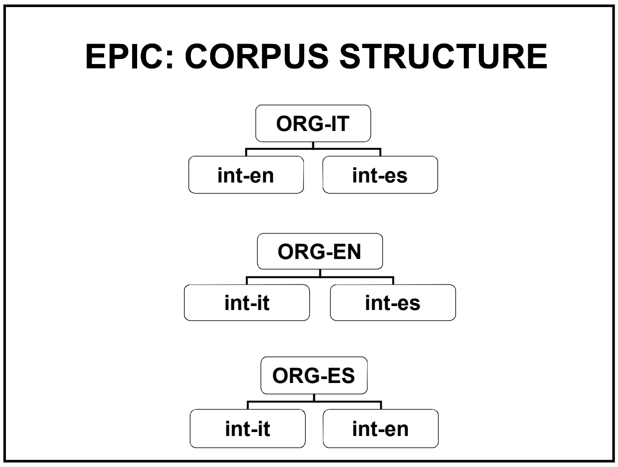
int-en | int-es

ORG-EN

int-it | int-es

ORG-ES

int-it | int-en



EPIC: CORPUS STRUCTURE

ORG-IT

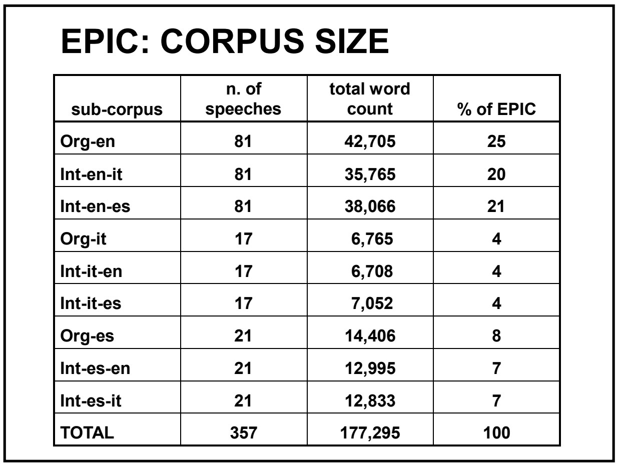
int-en | int-es

ORG-EN

int-it | int-es

ORG-ES

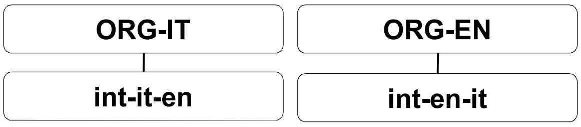
int-it | int-en



EPIC: CORPUS SIZE

sub-corpus	n. of speeches	total word count	% of EPIC
Org-en	81	42,705	25
Int-en-it	81	35,765	20
Int-en-es	81	38,066	21
Org-it	17	6,765	4
Int-it-en	17	6,708	4
Int-it-es	17	7,052	4
Org-es	21	14,406	8
Int-es-en	21	12,995	7
Int-es-it	21	12,833	7
TOTAL	357	177,295	100

DIRSI: CORPUS STRUCTURE


ORG-IT

int-it-en

ORG-EN

int-en-it

DIRSI: CORPUS SIZE

sub-corpus	No. of texts	No. of words	% 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
TOTAL	158	135,835	100

Interpreter	Conference	Floor time in DIRSI-C
IT-01	CFF4 - CFF5	54’ + 98’ (152’)
IT-02	CFF5	129’
IT-03	ELSA	78’
IT-04	ELSA	77’
UK-01	CFF4	125’
Total		565’ (9h 25’ x 2)

CORPUS DESIGN: representativeness

a sample which is maximally representative of the variety under examination

(McEnery & Wilson 2001, p. 30)

DATA COLLECTION:

- **EPIC**: Europe by Satellite (2004)
  EP website (since April 2006)
- **DIRSI**: field work

EPIC Multimedia Archive
(audio/video)

1. Collection: EbS channel video-recordings (140 VHS tapes – 5 part-sessions)

2. Digitisation: .mpeg (org) – Pinnacle Studio
.wav (int) – Cooledit

3. File editing: clips e.g. 10-02-04-m-005-org-en
10-02-04-m-005-int-en-it
10-02-04-m-005-int-en-es

DIRSI Multimedia Archive
(audio)

Collection: audio-recordings
(14 conferences, laptop + minirecorder)

File editing: .wav
Cooledit Pro 2.0 + Audacity (1.2.4)
clips e.g.
DIRSI-2006-05-20-VR-CFF4-001-org-it
DIRSI-2006-05-20-VR-CFF4-001-int-it-EN

DIRSI-C :
access to the target populations

Conference organizers + participants
PCO / agencies

Practisearcher **(self-analysis?)**
Institutional support
Networking
Inside champions

DATA COLLECTION:
informed consent

Conference organizers + participants
PCO / agencies

Broad vs. narrow scope (research only?)
No commercial purposes
Interpreters' anonymity
No publication of conference proceedings
Shared responsibility

DATA COLLECTION: recording (1)

2 communication flows:

floor (ST) + booth (TT)

External vs. internal access
On site technical staff
Conference hall audio system
Digital mini recorders
Laptop / Microphone in the booth

DATA COLLECTION: recording (2)

2 communication flows:

floor (ST) + booth (TT)

Researcher's kit
Beware of default settings (e.g. energy saving)
Formats (.wav vs. .mp3)
Back up copies
Data storage system (multimedia archive)

TRANSCRIPTION

- No international standard
- Spoken language features
- Machine-readable vs. User-friendly
- Alignment

Examples from EPIC: org int int

(date: 10-02-04-m
speech number: 015
language: en
type: org-en

duration: medium
timing: 137

text length: medium
number of words: 339

speed: medium
words per minute: 148

source text delivery: read

speaker: Parish, Neil
gender: M
country: United Kingdom
mother tongue: yes

political function: MEP
political group: PPE-DE

topic: Health
specific topic: Asian bird flu

comments: NA)

where is the ... trying to get the microphone working yeah // Commissioner Byrne I welcome very much your statement here here this morning // but I understand Sir you went to Thailand and were told that it wasn't avian flu but it was chicken cholera </chorela/> cholera // and it was only when you returned that you later found that it was actually chicken flu // you I welcome the ban you put on Thai chicken meat but of course you have not banned cooked meat or canned meat and this makes up about seventy-five per cent of our imports from Thailand // the Thai authorities have been found wanting in conp- with in- international compliance // they have hidden the Thai flu for maybe two months // if they protect if they if they protect their market fro- from a deadly disease surely they will protect it against minor infringements of EU standards // will they run the same risk with cooked meats // will we see nitrofuran residues and hormones coming in in this in this cooked meat // is the public health of the EU overridden by trade interests of EU exporters // I believe Thailand should be completely delisted from all poultry meat exports until they can prove they have the infrastructure and the integrity to apply EU standards all the time // can we also have a labelling of country of origin for principal ingredients in processed food because this is where most of this Thai and cooked meat comes in // ... how are we sure that meat cooked in Thailand has reached the correct temperature // maybe a slight risk how do we know // cooked meat cooked chicken will come into the UK frozen thawed then eaten with no further treatment from when the Thai cooked it // is this a risk to our consumers // finally the tariff the the Thai chicken meat cooked chicken meat is one seventh of that on frozen fresh chicken // why is this so Commissioner because this seems to be a loophole // and finally has the disease moved to the United States of America

(date: 10-02-04-m
speech number: 015
language: es
type: int-en-es

duration: medium
timing: 137

text length: short
number of words: 228

speed: low
words per minute: 100

source text delivery: read

speaker: interpreter
gender: F
country: NA
mother tongue: yes

political function: NA
political group: NA

topic: Health
specific topic: Asian bird flu

comments: NA)

señor Comisario celebro mucho su declaración aquí esta mañana // pero tengo entendido que usted estuvo en Tailandia y se le dijo que no era ehm pe- peste aviar gripe aviar sino que era cólera // pero ehm celebro que haya una prohibición de importar carne de ave de Tailandia pero no se ha prohibido la importación de ehm pa- partes de estas aves de carne cortada ... // ehm la digamos ya se conoce este caso de gripe de Tailandia desde hace un par de meses pero si se protege </protofe/> protege su carne protegemos digamos nuestros mercados no tendríamos que hacer lo mismo con carne cortada también que ehm ... la salud pública de la UE no se encuentra más bien dominada por los intereses económicos de los operadores // tendríamos que prohibir todas las importaciones de carne hasta que se dé la seguridad de que se aplican normas comunitarias // ehm podríamos tener un etiquetado del país de origen para los principales ingredientes en ehm alimentos procesados porque hay mucha carne tailandesa que entra en esos preparados // se está seguro que la los preparados han alcanzado las temperaturas adecuados y el pollo llega al Reino Unido ehm congelado y se come sin ninguna ehm tratamiento adicional que como se cocinó en Tailandia en su origen // ehm la carne tailande- de pollo tailandesa digamos ehm es ehm mucho más barata que la otra que carne que se puede comprar en otros que procede de otras destinos

TRANSCRIPTION CONVENTIONS

truncated words	propo	propo-
	pro posal	proposal </pro_posal/>
mispronounced words	Parlomento	Parlamento </Parlomento/>
pauses	filled	ehm
	empty	…
numbers	532	five hundred and thirty-two
percentages	4%	four per cent
dates	1997	nineteen ninety-nine
unintelligible		#
units	(syntax & intonation)	// (no punctuation signs)

TRANSCRIBING (EPIC)

.txt

“FIRST DRAFT”

ST: verbatim reports from EP website

TT: speech-recognition software (“shadowing”)

“FINAL DRAFT”

proofing and crosschecking

Tagging and indexing EPIC

English:	TreeTagger (Schmid 1994) Revised version of the Penn-Treebank tagset
Italian:	TreeTagger (36 tags)
Spanish:	FreeLing (Carreras et al. 2004)
[Spanish:	GRAMPAL] (Moreno & Guirao)

- Corpus WorkBench CWB-CQP (Christ 1994)

Alignment

- Sound - Text
- ST – TT
  - Tabular
  - Musical Score

Access and Distribution: EPIC

http://www.sslmitdev-online.sslmit.unibo.it/corpora/corpora.php

- **Automatic extraction of occurrences**
  - SIMPLE QUERY
  - ADVANCED QUERY
  - **Search parameters** (header)
- ELRA catalogue

  European Language Resources Association

Access and Distribution: DIRSI-C

- **Corpus Work Bench (CWB-CQP)**
- **LLI-UAM portal**
- http://drusila.lllf.uam.es/lab/
- http://cartago.lllf.uam.es/dir-si/
- **Flexible formats!**

References

Armstrong, Susan (1997) Corpus-based methods for NLP and translation studies. *Interpreting* 2/1-2, pp. 141-162.

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 M. Baker, Francis, G. & E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, pp. 233-250.

Baker, Mona (1995) Corpora in Translation Studies. An overview and suggestions for future research. *Target* 7/2, pp. 223-43.

Baker, Mona (1996) Corpus-based Translation Studies. The challenges that lie ahead. In Somers, H. (ed.) *Terminology, LSP and Translation*. Amsterdam/Philadelphia: John Benjamins, pp. 175-86.

Baker, Mona (1998) Réexplorer la langue de la traduction: une approche par corpus. *Meta* 43/4, pp. 480-485.

Baker, Mona (1998) The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4, pp. 281-298.

Shlesinger, Miriam (1998) Corpus-based Interpreting Studies as an offshoot of Corpus-based Translation Studies. *Meta* 43/4, pp. 486-493.

Laviosa, Sara (1998) 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* 35/1, pp. 6-27.

Setton, Robin (2011) Corpus-Based Interpreting Studies (CIS): Overview and Prospects. In A. Kruger, Walmach, K. & J. Munday (eds.) *Corpus-based Translation Studies: Research and Applications*. London/New York: Continuum.

Bowker, Lynne & Jennifer, Pearson (2002) *Working with Specialized Language. A practical guide to using corpora*. London/New York: Routledge.

Thompson, Paul (2005) *Spoken language corpora*. In Wynne, M. (ed.) Developing *Linguistic Corpora: A Guide to Good Practice*. Online: http://ota.ahds.ac.uk/documents/creating/dlc/chapter5.htm

Cencini, Marco (2000) Il Television Interpreting Corpus (TIC). *Proposta di codifica conforme alle norme TEI per trascrizioni di eventi di interpretazione in televisione*. Unpublished MA thesis, 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

McEnery, Tony & Andrew, Wilson (2001) *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing. September 1994*.

Carreras, Xavier; Chao, Isaac; Padró, Lluís & Muntsa, Padró (2004) Freeling: an open-source suite of language analyzers. In 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.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon: ELRA, pp. 239-242.

Christ, Oli (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System, *COMPLEX '94, Budapest*. Online (Corpus Work Bench info page): http://www3.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CorpusWorkbench.html

Moreno Sandoval, A. & Guirao, José María (2006). Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Kawaguchi, Yuji; Zaima, Susumu & Toshihiro, Takagaki (eds.) *Spoken Language Corpus and Linguistic Informatics*. Amsterdam/Philadelphia: John Benjamins, pp. 199-218.

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; 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.) *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins pp. 135-161.

You will find more references in the bibliography of these two books (they are written in Italian but they can be dowloaded for free and have many extra references):

Bendazzoli, C. (2010a) *Testi e contesti dell'interpretazione di conferenza: uno studio etnografico*. Bologna: Asterisco. [Open access: http://amsacta.unibo.it/2905/]

Bendazzoli, C. (2010b) *Corpora e interpretazione simultanea*. Bologna: Asterisco. [Open access: http://amsacta.unibo.it/2897/]