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**
- **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>
<thead>
<tr>
<th>Researcher</th>
<th>Languages</th>
<th>Event</th>
<th>Mode</th>
<th>Subjects</th>
<th>Length</th>
<th>Transcription published or available</th>
<th>Sound files availability</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oleron and Nanpon</td>
<td>EN, FR, DE, ES</td>
<td>UNESCO impromptu (non-tech discussion)</td>
<td>SI</td>
<td>pros</td>
<td>~7 minutes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Time lag, speed, fidelity</td>
</tr>
<tr>
<td>Dejean Le Feal 1978</td>
<td>FR&gt;DE</td>
<td>Various speeches</td>
<td>SI</td>
<td>pros</td>
<td>~40 hours'</td>
<td>77 pages, speakers and interpreters</td>
<td>Text on microfiches, AIIC*</td>
<td>Recited vs. impromptu input (école du sens)</td>
</tr>
<tr>
<td>Chernov 1978</td>
<td>EN, FR, ES, RU</td>
<td>UN 1968</td>
<td>SI</td>
<td>pros</td>
<td>~40 hours'</td>
<td>'Parallel transcripts' extracts published</td>
<td>Probably N. A.</td>
<td>Illustrate theory (redundancy &amp; prediction) (ditto)</td>
</tr>
<tr>
<td></td>
<td>EN&gt;RU, ES, FR</td>
<td>1978 UN satellite interpreting experiment</td>
<td>SI</td>
<td>pros</td>
<td>~40 hours'</td>
<td>(ditto)</td>
<td>Probably no longer available</td>
<td>(ditto)</td>
</tr>
<tr>
<td>Lederer 1981</td>
<td>DE&gt;FR</td>
<td>Railway Consortium and lab (2nd versions)</td>
<td>SI</td>
<td>2 pros</td>
<td>3 hours taped (original + 2 interpreters)</td>
<td>63 minutes original DE, FR; some extracts synchronized (interlinear)</td>
<td>Illustrate theory (école du sens)</td>
<td></td>
</tr>
<tr>
<td>Shlesinger 1989</td>
<td>HEB&lt;EN</td>
<td>Courtroom testimony</td>
<td>SI</td>
<td>pros</td>
<td>4 hours</td>
<td>50 pp SI (2 events), fluent text speaker + interpreter, some extracts synchronized (interlinear)</td>
<td>No longer available</td>
<td>Fidelity examples (école du sens)</td>
</tr>
<tr>
<td></td>
<td>FR&lt;EN</td>
<td>Extracts from 2 meetings 1986-88</td>
<td>SI</td>
<td>pros</td>
<td>~4 hours</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Researcher</td>
<td>Languages</td>
<td>Event</td>
<td>Mode</td>
<td>Subjects</td>
<td>Length</td>
<td>Transcription published or available</td>
<td>Sound files availability</td>
<td>Analysis</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-------</td>
<td>------</td>
<td>----------</td>
<td>--------</td>
<td>-----------------------------------</td>
<td>-------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Pöchhacker 1994</td>
<td>EN&lt;&gt;DE, FR&gt;DE</td>
<td>Vienna small business conference <em>ICS</em></td>
<td>SI</td>
<td>pros</td>
<td>14 hours (original + interpreter)</td>
<td>Available as vol. 3 of doctoral diss. from U. of Vienna library; parts published in Pöchhacker 94</td>
<td>Tapes available from author</td>
<td>Intersexuality, situational &amp; delivery factors (speed, slips/shifts, hesitation EN&lt;&gt;DE)</td>
</tr>
<tr>
<td>Kalina 1998</td>
<td>EN&lt;&gt;DE&lt;&gt;FR</td>
<td>Bertelt; 1989 public lecture (anti nuclear)</td>
<td>SI</td>
<td>6 students</td>
<td>70 minutes</td>
<td>5 versions tiered. 5s per line.</td>
<td>Consul, loan at Heidberg</td>
<td>Choose examples, identify errors, strategies</td>
</tr>
<tr>
<td>DE, FR, EN</td>
<td>Würzburg; 1992 law symposium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(dito)</td>
<td>(dito)</td>
</tr>
<tr>
<td>Setton 1997, 99</td>
<td>DE&gt;EN</td>
<td>Excerpts from Kalina Würzburg corpus</td>
<td>SI</td>
<td>6 pros (2 per booth)</td>
<td>14/30 minutes microanalysis</td>
<td>I-track audio (simulation: 2 track)</td>
<td>Available from author</td>
<td>Linguistic, cognitive-pragmatic analysis for process modelling</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(dito)</td>
<td>(dito)</td>
</tr>
<tr>
<td>Wallmach 2000</td>
<td>EN, Zulu, Afrikaans, Sepedi</td>
<td>Parliament speeches Gauteng (S. Africa) Provincial legislature</td>
<td>SI</td>
<td>1 pro + 2 in mock</td>
<td>14 pros</td>
<td>Pilot: 6–8 hours EN, Afrikaans, Zulu</td>
<td>Pilot material transcribed (6–8 hrs)</td>
<td>110 hours on tape (3 languages)</td>
</tr>
<tr>
<td>Cencini 2000</td>
<td>EN&lt;IT</td>
<td>Television Interpreting Corpus (TIC), Comparable EN(18) and IT source speeches;</td>
<td>CI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TEI Standard (computer query-able)</td>
</tr>
<tr>
<td>Fumagalli 1999–2000</td>
<td>EN&gt;IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Features of ‘interpretee’ MultiConcord-Parallel Concordancer</td>
<td></td>
</tr>
<tr>
<td>Diriker 2001</td>
<td>EN&lt;TR</td>
<td>Conference on Metaphysics and Politics (2 days)</td>
<td>SI</td>
<td>pros</td>
<td>150 pp. transcripts</td>
<td>Available as Annex to Ph.D.</td>
<td>Bogazici Univ. Library</td>
<td>Interaction and sociological</td>
</tr>
<tr>
<td>Source</td>
<td>Language</td>
<td>Type</td>
<td>SI/CI</td>
<td>Duration</td>
<td>Output</td>
<td>Website</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>----------</td>
<td>--------</td>
<td>---------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Vuorikoski 2004</td>
<td>mostly DE, EN &gt;FI, SW</td>
<td>European Parliament debates</td>
<td>SI</td>
<td>Ca. 70 pros</td>
<td>120 speeches, 65 analysed</td>
<td>Selected transcripts appended to PhD</td>
<td>Difficulty, quality in rhetorical (political) speech genre</td>
<td></td>
</tr>
<tr>
<td>Beacons 2007</td>
<td></td>
<td>European Parliament debates via EBS</td>
<td>SI</td>
<td>7 hours</td>
<td></td>
<td></td>
<td>Text-discourse-ideology link through cohesive devices</td>
<td></td>
</tr>
<tr>
<td>Monacelli 2005, 2009</td>
<td>IT&gt;EN</td>
<td>10 speeches from 4 events</td>
<td>SI</td>
<td>10 pros</td>
<td>2 hours, interpretation</td>
<td>Selected transcripts in PhD</td>
<td>Interactive politeness, 'face', through deixis, mood, transitivity</td>
<td></td>
</tr>
<tr>
<td>ECIS Group (Univ. of Granada)</td>
<td>EN&gt;DE, ES, FR, DE, FR&gt;ES</td>
<td>European Parliament debates via EBS</td>
<td>SI</td>
<td>pros</td>
<td>43 speeches, 73 interpretation</td>
<td>Linked files in MS Access, tool developed for multivariate visualization</td>
<td>Available from authors</td>
<td></td>
</tr>
<tr>
<td>CIAIR (Nagoya University)</td>
<td>EN&gt;JA</td>
<td>Monologue SI (simple lectures) and travel dialogues</td>
<td>SI and CI</td>
<td>SI: 4 pros per speech</td>
<td>182 hrs (1m words)</td>
<td>Web access/interface</td>
<td>Quality minimizers and maximizers, verbal and non-verbal</td>
<td></td>
</tr>
<tr>
<td>EPIC**; Bologna-Forli group</td>
<td>EN-IT-ES (9 sub-corpora)</td>
<td>European Parliament sessions</td>
<td>SI</td>
<td>Multiple pros</td>
<td>Open-ended; 280 hrs usable in archive; ca. 280,000 words transcribed as of mid-2010</td>
<td>Available for query online (POS-tagged and indexed database)</td>
<td>Strategies, tag, etc. for machine SI research. Own software.</td>
<td></td>
</tr>
</tbody>
</table>

As of 2010: directionality, lexis, disfluencies; Ongoing (+ grad theses). Semi-automatic transcription technique
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

Methodological challenges

1. Corpus design
   - >> corpus structure
   - >> representativeness

2. Data collection
   - >> accessibility
   - >> informed consent
   - >> recording

CORPUS DESIGN: structure

- ORG-IT
- ORG-EN
- int-it-en
- int-en-it
- int-en
- int-es
- int-it
- int-es
- int-it
- int-en

Settings:

Sources:
EPIC: CORPUS STRUCTURE

EPIC: CORPUS SIZE

<table>
<thead>
<tr>
<th>sub-corpus</th>
<th>n. of speeches</th>
<th>total word count</th>
<th>% of EPIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Org-en</td>
<td>81</td>
<td>42,705</td>
<td>25</td>
</tr>
<tr>
<td>Int-en-it</td>
<td>81</td>
<td>35,765</td>
<td>20</td>
</tr>
<tr>
<td>Int-en-es</td>
<td>81</td>
<td>38,066</td>
<td>21</td>
</tr>
<tr>
<td>Org-it</td>
<td>17</td>
<td>6,765</td>
<td>4</td>
</tr>
<tr>
<td>Int-it-en</td>
<td>17</td>
<td>6,708</td>
<td>4</td>
</tr>
<tr>
<td>Int-it-es</td>
<td>17</td>
<td>7,052</td>
<td>4</td>
</tr>
<tr>
<td>Org-es</td>
<td>21</td>
<td>14,406</td>
<td>8</td>
</tr>
<tr>
<td>Int-es-en</td>
<td>21</td>
<td>12,995</td>
<td>7</td>
</tr>
<tr>
<td>Int-es-it</td>
<td>21</td>
<td>12,833</td>
<td>7</td>
</tr>
<tr>
<td>TOTAL</td>
<td>357</td>
<td>177,295</td>
<td>100</td>
</tr>
</tbody>
</table>
DIRSI: CORPUS STRUCTURE

ORG-IT
  int-it-en

ORG-EN
  int-en-it

DIRSI: CORPUS SIZE

<table>
<thead>
<tr>
<th>sub-corpus</th>
<th>No. of texts</th>
<th>No. of words</th>
<th>% DIRSI-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORG-IT</td>
<td>63</td>
<td>33,412</td>
<td>24.6</td>
</tr>
<tr>
<td>INT-IT-EN</td>
<td>63</td>
<td>31,510</td>
<td>23.2</td>
</tr>
<tr>
<td>ORG-EN</td>
<td>16</td>
<td>37,249</td>
<td>27.4</td>
</tr>
<tr>
<td>INT-EN-IT</td>
<td>16</td>
<td>33,664</td>
<td>24.8</td>
</tr>
<tr>
<td>TOTAL</td>
<td>158</td>
<td>135,835</td>
<td>100</td>
</tr>
</tbody>
</table>

Interpreter Conference Floor time in DIRSI-C

<table>
<thead>
<tr>
<th>Interpreter</th>
<th>Conference</th>
<th>Floor time in DIRSI-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT-01</td>
<td>CFF4 · CFF5</td>
<td>54' + 98'' (152'')</td>
</tr>
<tr>
<td>IT-02</td>
<td>CFF5</td>
<td>129'</td>
</tr>
<tr>
<td>IT-03</td>
<td>ELSA</td>
<td>78'</td>
</tr>
<tr>
<td>IT-04</td>
<td>ELSA</td>
<td>77'</td>
</tr>
<tr>
<td>UK-01</td>
<td>CFF4</td>
<td>125'</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>565'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9h 25' x 2)</td>
</tr>
</tbody>
</table>
CORPUS DESIGN: representativeness

a sample which is maximally representative of the variety under examination

(McEnery & Wilson 2001, p. 30)

DATA COLLECTION:

  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)
clicks 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
<table>
<thead>
<tr>
<th>date: 10-02-04-m</th>
<th>date: 10-02-04-m</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech number: 015</td>
<td>speech number: 015</td>
</tr>
<tr>
<td>language: en</td>
<td>language: es</td>
</tr>
<tr>
<td>type: org-en</td>
<td>type: int-en-es</td>
</tr>
<tr>
<td>duration: medium</td>
<td>duration: medium</td>
</tr>
<tr>
<td>timing: 137</td>
<td>timing: 137</td>
</tr>
<tr>
<td>text length: medium</td>
<td>text length: short</td>
</tr>
<tr>
<td>number of words: 339</td>
<td>number of words: 228</td>
</tr>
<tr>
<td>speed: medium</td>
<td>speed: low</td>
</tr>
<tr>
<td>words per minute: 148</td>
<td>words per minute: 100</td>
</tr>
<tr>
<td>source text delivery: read</td>
<td>source text delivery: read</td>
</tr>
<tr>
<td>speaker: Parish, Neil</td>
<td>speaker: interpreter</td>
</tr>
<tr>
<td>gender: M</td>
<td>gender: F</td>
</tr>
<tr>
<td>country: United Kingdom</td>
<td>country: NA</td>
</tr>
<tr>
<td>mother tongue: yes</td>
<td>mother tongue: yes</td>
</tr>
<tr>
<td>political function: MEP</td>
<td>political function: NA</td>
</tr>
<tr>
<td>political group: PPE-DE</td>
<td>political group: NA</td>
</tr>
<tr>
<td>topic: Health</td>
<td>topic: Health</td>
</tr>
<tr>
<td>specific topic: Asian bird flu</td>
<td>specific topic: Asian bird flu</td>
</tr>
<tr>
<td>comments: NA</td>
<td>comments: NA</td>
</tr>
</tbody>
</table>

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 </cholera> 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 comp- 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

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 tendriamos 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

<table>
<thead>
<tr>
<th>truncated words</th>
<th>pro proosal</th>
<th>proproposal</th>
<th>proposal &lt;/pro_posal&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>mispronounced</td>
<td>Parlamento</td>
<td>Parlamento</td>
<td>Parlamento &lt;/Parlomento&gt;</td>
</tr>
<tr>
<td>words</td>
<td>filled</td>
<td>empty</td>
<td>ehm ...</td>
</tr>
<tr>
<td>pauses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>numbers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>percentages</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>units</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.sslmit-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.lifl.uam.es/lab/
  - http://cartago.lifl.uam.es/dir-si/

- Flexible formats!
References


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


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