From EPIC to EPTIC - building and using an intermodal corpus of translated and interpreted texts

Silvia Bernardini[1], Adriano Ferraresi[1] and Maja Miličević[2]

[1]University of Bologna, [2]University of Belgrade

New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies

SLE 2013 - Split, 18 September 2013

Overview

1. Background
2. Building EPTIC
   - Corpus composition
   - Corpus mark-up
3. Using EPTIC
   - Case study on lexical simplification
   - A more qualitative perspective
4. Conclusion

Parallel and comparable corpora

Translated and interpreted texts are typically studied in relation to:

- Their source texts (parallel corpora)
- Comparable original texts (comparable corpora)
    - Translation → written non-translated production
    - Interpreting → oral non-translated production

Intermodal corpora

translation scholars can learn about the process and product of (written) translation by finding out more about interpreting – and interpreting scholars can infer about this high-pressure form of translation by observing the slower, more readily observable process and product of (written) translation

(Shlesinger and Ordan 2012: 44)

Intermodal corpora

Corpora comprising both translated and interpreted texts

- **Kajzer-Wietrzny (2012)**
  - a monolingual comparable and intermodal corpus based on the European Parliament plenary sessions, containing interpreted and translated texts (French/Spanish/German/Dutch > English), as well as texts originally produced in English
- **Shlesinger (2009), Shlesinger and Ordan (2012)**
  - a small-scale, monolingual comparable and intermodal corpus comprising translational and interpretational outputs of the same source text by six professional translators/interpreters (English > Hebrew, within-subject, experimental data)
  - a monolingual comparable and intermodal corpus comprising translational and interpretational outputs and spontaneous speeches in the academic domain (English > Hebrew, authentic data)

A new intermodal corpus

EPTIC < European Parliament Translation and Interpreting Corpus

→ An extension of **EPIC** (European Parliament Interpreting Corpus)

From EPIC to EPTIC

EPIC is a trilingual (English ↔ Italian ↔ Spanish) corpus of European Parliament speeches and their corresponding interpretations (Sandrelli and Bendazzoli 2005, Bendazzoli 2010)

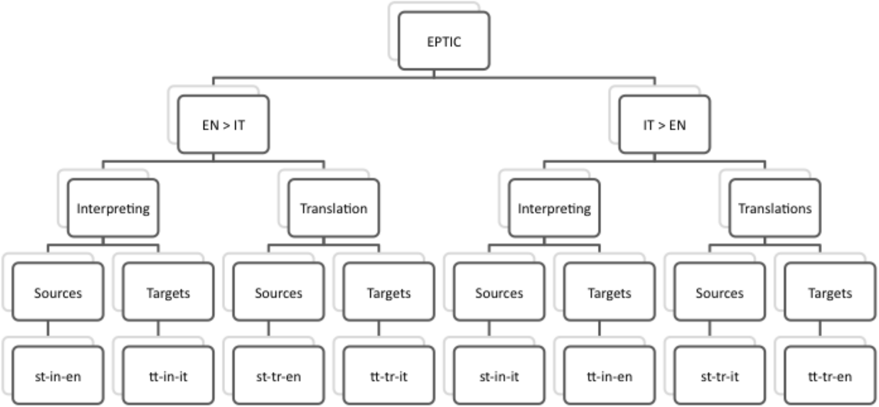
EPIC > EPTIC

- Transcripts of interpreted talks and their source texts were taken from EPIC [✓ English, ✓ Italian, ✗ Spanish]
- The revised source texts and their (independently produced) translations were obtained from the European Parliament website

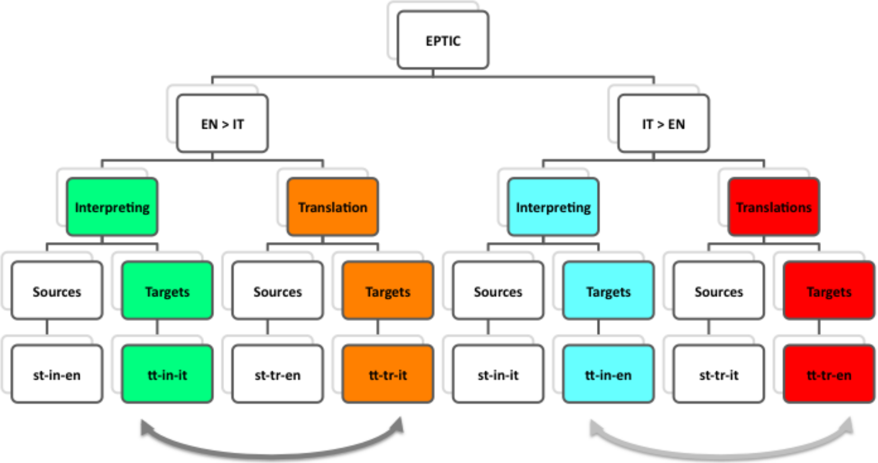
As a result, EPTIC is:

- A bilingual, bidirectional corpus (English ↔ Italian)
- An intermodal, comparable and parallel corpus comprising simultaneous interpretations paired with their source texts + corresponding translations and source texts (8 components)

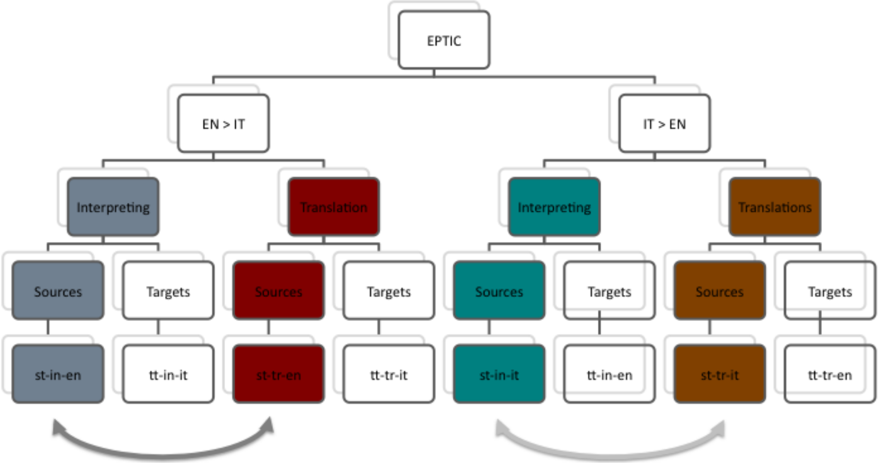
Corpus structure



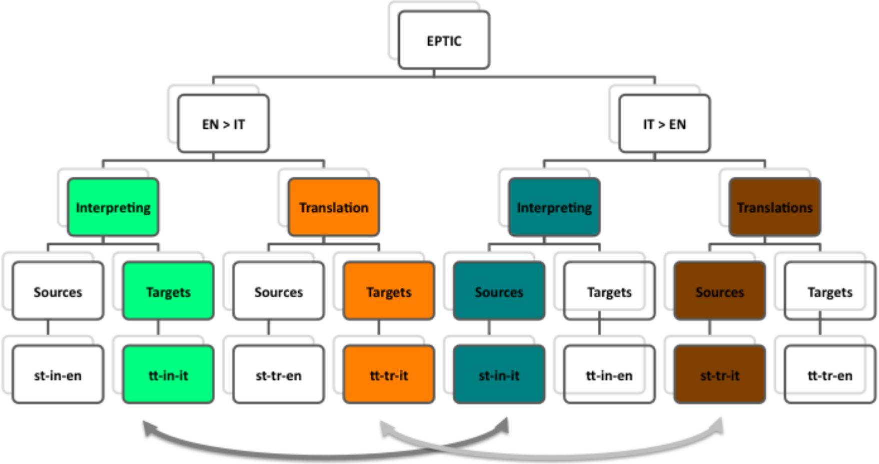
Intermodal corpora (targets)



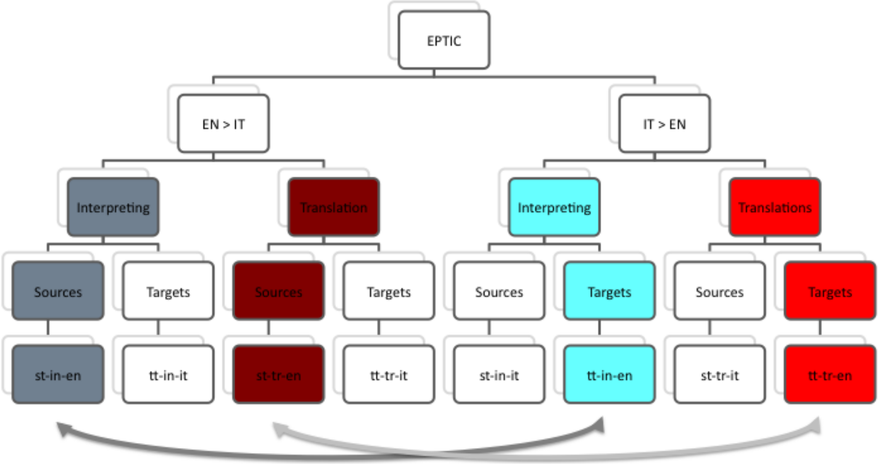
Intermodal corpora (sources)



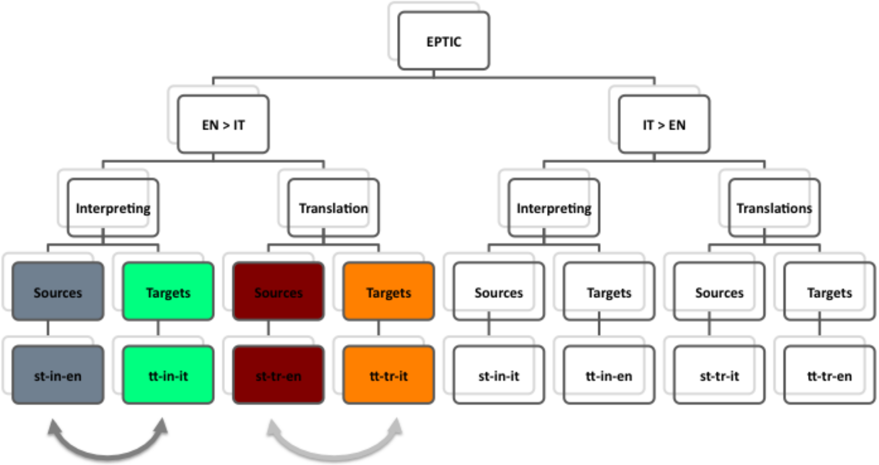
Comparable corpora (Italian)



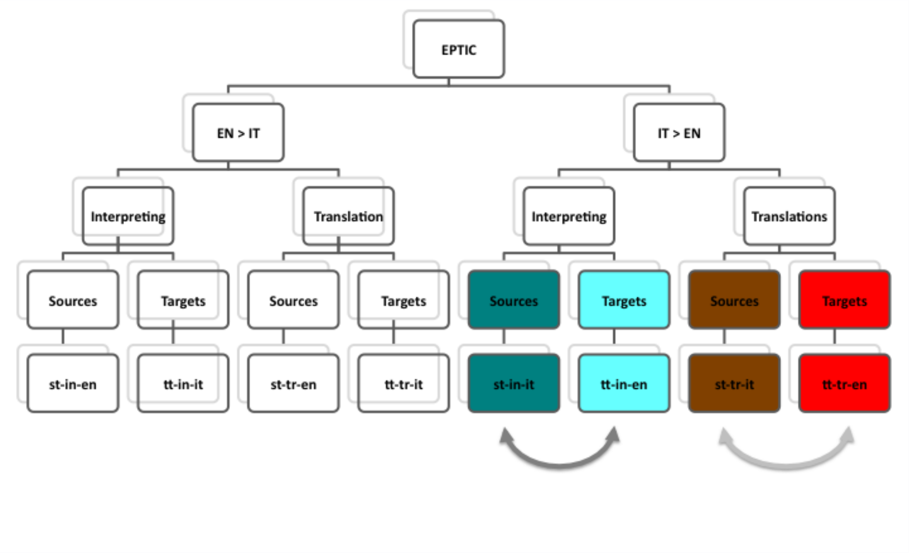
Comparable corpora (English)



Parallel corpora (EN > IT)



Parallel corpora (IT > EN)



Size

Subcorpus	N. of texts	Total word count*	% of EPTIC
st-in-en	81	41,869	23.91
st-tr-en	81	36,685	20.95
tt-in-it	81	33,675	19.23
tt-tr-it	81	36,876	21.06
Subtotal	324	149,105	85.14
st-in-it	17	6,387	3.65
st-tr-it	17	6,234	3.56
tt-in-en	17	6,577	3.76
tt-tr-en	17	6,819	3.89
Subtotal	68	26,017	14.86
Total	**392**	**175,122**	**100.00**

*Truncated words in interpreted texts are omitted from the count.

Text preprocessing

Existing plain text files with metadata headers were taken from EPIC

Corresponding files were created for translation source and target texts, with relevant metadata from EPIC's files

Metadata

Available metadata include:

- Speaker information (identity, gender, country, political affiliation, L1)
- Interpreter information (gender, L1)
- Speech delivery type (read, impromptu or mixed)
- Speech length (short, medium or long)
- Speech topic (general and specific)

Example - target text header (int.):

```
#text id=2 date=10-02-04-m speech=006
length=medium duration=medium
delivery=mixed topic=Health
topicspec=Asian-bird-flu speaker
name=Jackson-Caroline-F. gender=F
country=United-Kingdom native=y
politfunc=MEP politgroup=PPE-DE
interpreter gender=F native=y
```

Example - source text header (int.):

```
#text id=2 date=10-02-04-m speech=006
length=medium duration=medium
delivery=mixed topic=Health
topicspec=Asian-bird-flu speaker
name=Jackson-Caroline-F. gender=F
country= United-Kingdom native=y
politfunc=MEP politgroup=PPE-DE
```

Linguistic mark-up and alignment

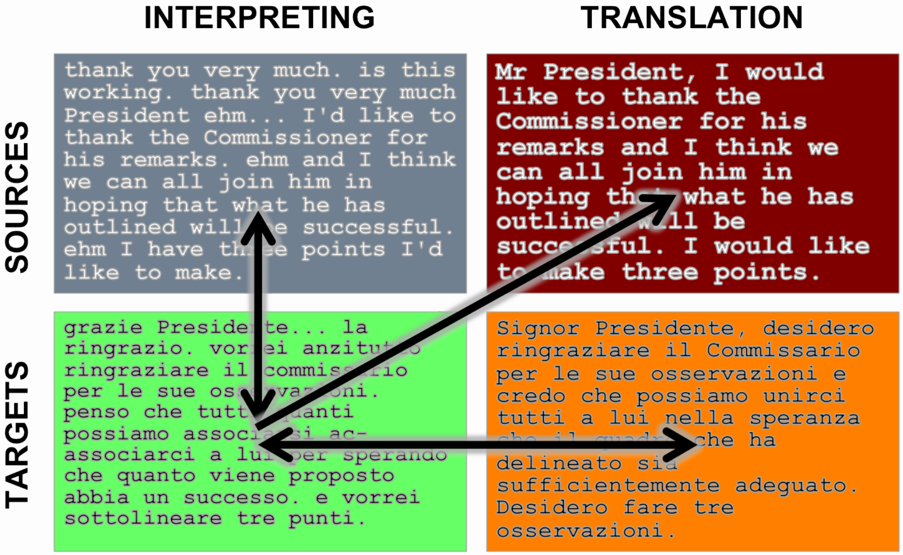
Linguistic mark-up was added independently of EPIC:

- Part-of-speech tagging and lemmatisation
  → TreeTagger
- Indexing
  → Corpus WorkBench

Sentence-level alignment was performed for:

- Parallel (source-target) pairs
- Intermodal (translation-interpretation) pairs

E.g., the EN > IT sub-corpus



INTERPRETING

TRANSLATION

SOURCES

TARGETS

thank you very much. is this working. thank you very much President ehm... I'd like to thank the Commissioner for his remarks. ehm and I think we can all join him in hoping that what he has outlined will be successful. ehm I have three points I'd like to make.

Mr President, I would like to thank the Commissioner for his remarks and I think we can all join him in hoping that what he has outlined will be successful. I would like to make three points.

grazie Presidente... la ringrazio. vorrei anzitutto ringraziare il commissario per le sue osservazioni. penso che tutti quanti possiamo associarsi ac- associarci a lui per sperando che quanto viene proposto abbia un successo. e vorrei sottolineare tre punti.

Signor Presidente, desidero ringraziare il Commissario per le sue osservazioni e credo che possiamo unirci tutti a lui nella speranza che il quadro che ha delineato sia sufficientemente adeguato. Desidero fare tre osservazioni.

Lexical simplification

Hypothesized to be a translation/interpretation universal, but previous studies report mixed results

- Texts translated into English are lexically simpler than comparable non-translated English texts (Laviosa 1998)
- Texts interpreted into English are less lexically simple than native English speeches than comparable non-interpreted English texts (Kajzer-Wietrzny 2012)
- Lexical simplification in interpreted texts appears to depend on the language combination (Sandrelli and Bendazzoli 2005)
- **Texts interpreted from English into Hebrew are simpler than translated texts** (Shlesinger and Ordan 2012)

Method (1a)

Measures of lexical simplification (following Laviosa 1998):

- **Lexical density**
  - Proportion of lexical to function words
  - Calculated as lexical words/total running words, i.e. (total running words - function words)/total running words
- **List heads**
  - The percentage of corpora covered by the first hundred words of their frequency lists
- **Core vocabulary**
  - The proportion of high frequency words to low frequency words calculated with reference to a list of the 200 most frequent words in English/Italian (extracted from ukWaC and itWaC respectively)

Method (1b)

Other measures looked at ([mostly] following Laviosa 1998):

- **Type-token ratio**
  - Proportion of unique words to the total number of words
- **Sentence length**
  - Mean number of words in a sentence
- **Variance**
  - In lexical density, type-token ratio and sentence length

These measures will not be discussed further (→ inconclusive results)

Method (2)

The focus is on **intermodal comparisons**:

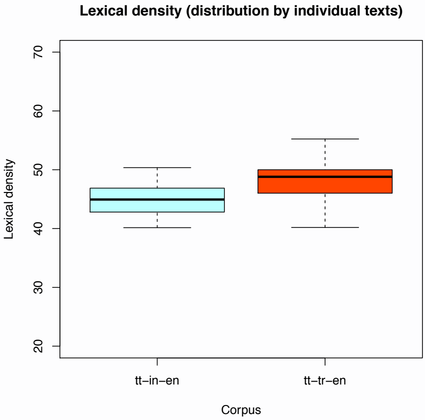
- Interpreted target texts are compared to translated target texts (tt-in-en *vs.* tt-tr-en / tt-in-it *vs.* tt-tr-it)
- Interpreting and translation source texts constitute a control comparison (st-in-it *vs.* st-tr-it / st-in-en *vs.* st-tr-en)

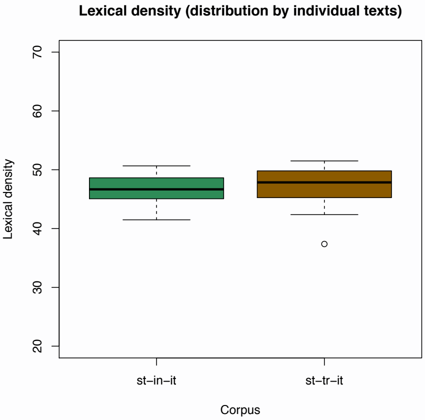
Method (3)

Statistical tests:

- **Lexical density**
  - Calculated by single texts
  - Mann-Whitney tests
- **List heads**
  - Calculated for each sub-corpus as a whole
  - Chi-square tests
- **Core vocabulary**
  - Calculated for each sub-corpus as a whole
  - Chi-square tests

Results: Lexical density in interpreted vs. translated English

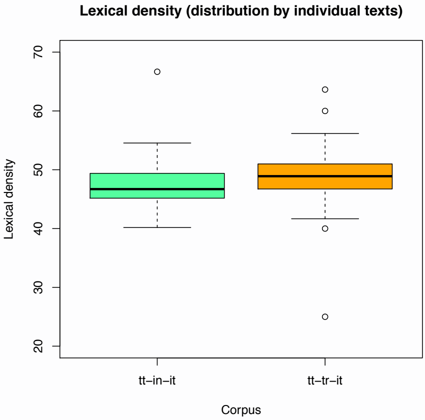


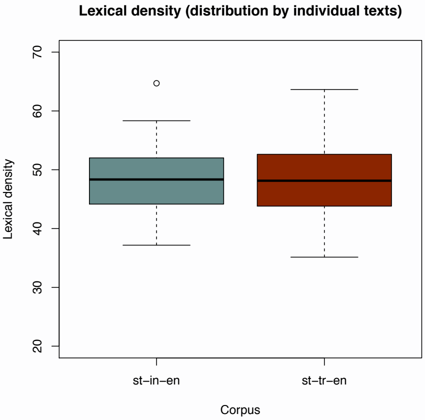


tt-in-en	tt-tr-en
44.94%	48.80%
W=82, $p<0.05$	

st-in-it	st-tr-it
46.67%	47.85%
ns	

Results: Lexical density in interpreted vs. translated Italian

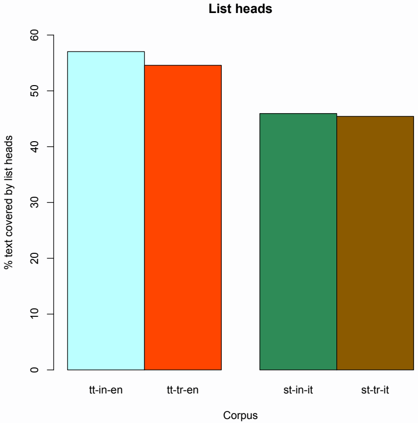




tt-in-it	tt-tr-it
46.71%	48.90%
W=2275.5, p<0.001	

st-in-en	st-tr-en
48.34%	48.14%
ns	

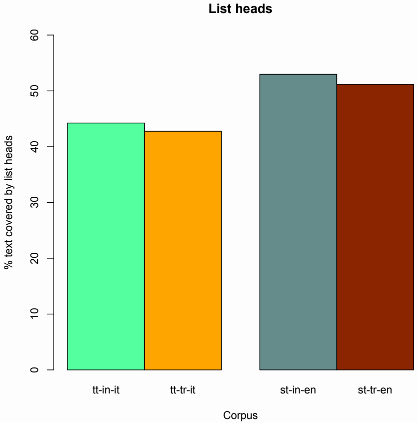
Results: List heads in interpreted vs. translated English



tt-in-en	**tt-tr-en**
57.03%	54.58%
$\chi^2(1) = 8.0437$, $p<0.01$	

st-in-it	**st-tr-it**
45.94%	45.43%
ns	

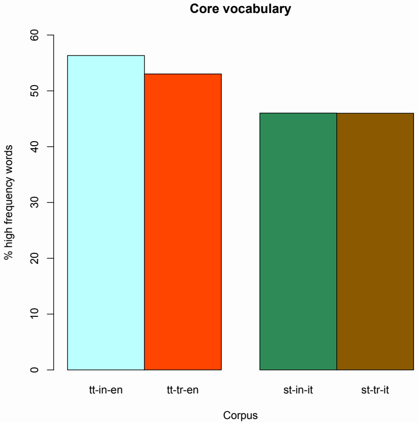
Results: List heads in interpreted vs. translated Italian



tt-in-it	tt-tr-it
44.24%	42.76%
$\chi^2(1) = 15.7206$, $p < 0.001$	

st-in-en	st-tr-en
52.97%	51.14%
$\chi^2(1) = 26.2893$, $p < 0.001$	

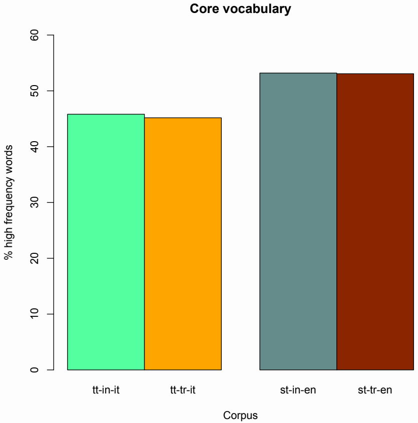
Results: Core vocab. in interpreted vs. translated English



tt-in-en	tt-tr-en
56.33%	53.03%
$\chi^2(1) = 14.6167$, p<0.001	

st-in-it	st-tr-it
46.02%	45.99%
ns	

Results: Core vocab. in interpreted vs. translated Italian



tt-in-it	tt-tr-it
45.81%	45.17%
ns	

st-in-en	st-tr-en
53.19%	53.07%
ns	

Summing up

Lexical simplification in interpreted vs. translated texts:

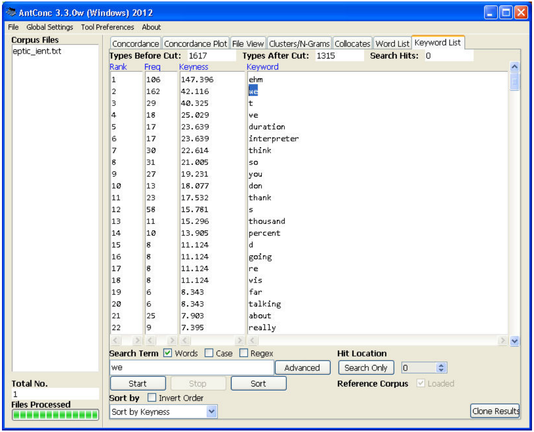
- Lexical density lower in interpreted English and Italian
- List heads cover higher percentages of interpreted English texts (no valid evidence for Italian)
- Core vocabulary covers higher percentages of interpreted English (but not Italian) texts

→ **Some evidence of interpreted texts being lexically simpler than translated texts in both directions**

→**Tendency stronger in IT > EN direction**

From keyword lists...

In the keyword list for interpreted English target texts (with translated texts as reference) the second position is occupied by *we*



Corpus Files
eptic_ient.txt

Types Before Cut: 1617 Types After Cut: 1315 Search Hits: 0

Rank	Freq	Keyness	Keyword
1	106	147.396	ehm
2	162	42.116	we
3	29	40.325	t
4	18	25.029	ve
5	17	23.639	duration
6	17	23.639	interpreter
7	30	22.614	think
8	31	21.005	so
9	27	19.231	you
10	13	18.077	don
11	23	17.532	thank
12	58	15.781	s
13	11	15.296	thousand
14	10	13.905	percent
15	8	11.124	d
16	8	11.124	going
17	8	11.124	re
18	8	11.124	vis
19	6	8.343	far
20	6	8.343	talking
21	25	7.903	about
22	9	7.395	really

Search Term ☑ Words ☐ Case ☐ Regex
we

Hit Location 0

Reference Corpus ☑ Loaded

Sort by ☐ Invert Order
Sort by Keyness

Total No.
1
Files Processed

… to parallel concordances (1)

Interpreted texts tend to be more personal, while translations seem to use more passive and impersonal forms

- **tt-in-en:** Do n't do n't does n't the Commission think that we should deal with this in the WTO // We should be talking about links between globalisation of trade and adverse health ehm e- epidemics
- **tt-tr-en:** Does the Commission not believe that, on this subject, a think-tank should be set up , within the actual context of the WTO , concerning the relationship between globalisation and health problems?
- **st-in-it:** La Commissione non ritiene che su questo tema vada inserito proprio in ambito WTO un tavolo di riflessione sul rapporto globalizzazione problemi sanitari
- **st-tr-it:** Non ritiene la Commissione che su questo tema vada inserito, proprio in ambito OMC, un tavolo di riflessione sul rapporto globalizzazione/problemi sanitari?

... to parallel concordances (2)

Interpreted texts tend to contain more verbal forms, while translations tend to prefer nominal forms (cf. also Shlesinger and Ordan (2012))

- **tt-in-en:** `But I think that we must be more realistic in our assessment of what Europe 's economic situation really is //`
- **tt-tr-en:** `... but with a more objectively realistic vision of the European economic framework.`
- **st-in-it:** `... ma con una visione oggettivamente più realistica del quadro q- economico europeo`
- **st-tr-it:** `... ma con una visione oggettivamente più realistica del quadro economico europeo .`

Concluding remarks: EPTIC

- **A novel source of ecologically valid data about different modes of translation**
- **It allows multiple comparisons: parallel, as well as comparable and intermodal**
- **Experimental version designed as an extension of EPIC, but more (recent) texts (and languages) are available...**

Thank you!

silvia@sslmit.unibo.it

adriano@sslmit.unibo.it

m.milicevic@fil.bg.ac.rs

Acknowledgements

Many thanks to the EPIC team for kindly providing their materials

References

Bendazzoli, C. (2010). *Corpora e interpretazione simultanea*. Bologna: Asterisco.

Kajzer-Wietrzny, M. (2012). *Interpreting universals and interpreting style*. Ph. D. thesis, Adam Mickiewicz University.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta 43*, 557–570.

Sandrelli, A. and C. Bendazzoli (2005). Lexical patterns in simultaneous interpreting: a preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series 1*.

Shlesinger, M. (2009). Towards a definition of *Interpretese*: An intermodal, corpus-based study. In G. Hansen, A. Chesterman, and H. Gerzymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile*, pp. 237–253. Amsterdam: John Benjamins.

Shlesinger, M. and N. Ordan (2012). More *spoken* or more *translated*? Exploring a known unknown of simultaneous interpreting. *Target 24*, 43–60.