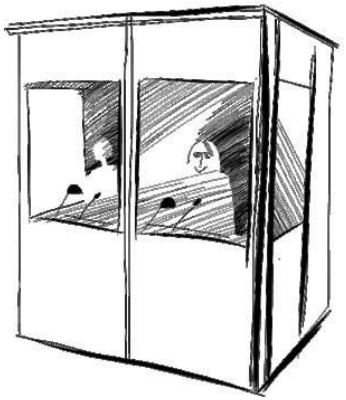
EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus



Heike Przybyl, Ekaterina Lapshinova-Koltunski,
Katrin Menzel, Stefan Fischer, Elke Teich

IDeaL
SFB 1102

UNIVERSITÄT
DES
SAARLANDES

EPIC UdS: Original vs interpreted

German original:
wir wollten / Kleinstunternehmen / und dabei sprechen wir von Unternehmen / die extrem klein sind / also mit ganz wenigen Angestellten minimalen Umsatzzahlen minimalen / Gewinnzahlen die im Grunde nur im regionalen Bereich vor Ort im lokalen Bereich tätig sind / der kleine Bäckermeister der kleine Malermeister / die wollen wir von den Bilanzverpflichtungen / befreien /

Interpreted into English:
we wanted to look at micro entities and that means / entities which are really very small with very few people working for them / minimum turnover / euh amoun/ minimum / profit amount / which are very / locally active / j/ just a small / hm / baker or painter and decorator we want to reduce the administrative burden on those companies

EPIC UdS: Original vs interpreted

German original:
wir wollten / Kleinstunternehmen / und dabei sprechen wir von Unternehmen / die extrem klein sind / also mit ganz wenigen Angestellten minimalen Umsatzzahlen minimalen / Gewinnzahlen die im Grunde nur im regionalen Bereich vor Ort im lokalen Bereich tätig sind / der kleine Bäckermeister der kleine Malermeister / die wollen wir von den Bilanzverpflichtungen / befreien /

Interpreted into English:
we wanted to look at micro entities and that means / entities which are really very small with very few people working for them / minimum turnover / euh amoun/ minimum / profit amount / which are very / locally active / j/ just a small / hm / baker or painter and decorator we want to reduce the administrative burden on those companies

EPIC UdS: Original vs interpreted

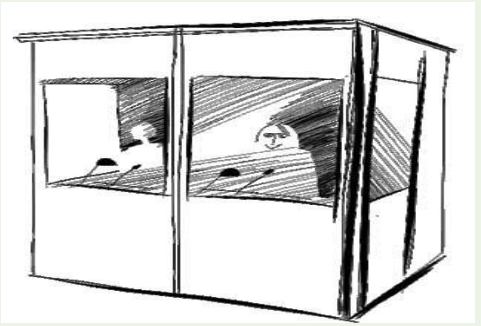
German original:
wir wollten / Kleinstunternehmen / und dabei sprechen wir von Unternehmen / die extrem klein sind / also mit ganz wenigen Angestellten minimalen Umsatzzahlen minimalen / Gewinnzahlen die im Grunde nur im regionalen Bereich vor Ort im lokalen Bereich tätig sind / der kleine Bäckermeister der kleine Malermeister / die wollen wir von den Bilanzverpflichtungen / befreien /

Interpreted into English:
we wanted to look at micro entities and that means / entities which are really very small with very few people working for them / minimum turnover / euh amoun/ minimum / profit amount / which are very / locally active / j/ just a small / hm / baker or painter and decorator we want to reduce the administrative burden on those companies

EPIC family of corpora

European Parliament Interpreting Corpora:

- **EPIC** – IT, EN, ES (Russo et al. 2005)
- **EPICG** – EN, FR, NL, ES (Defrancq et al. 2015)
- **TIC** – EN (Kajzer-Wietrzny 2012)
- PINC – PL, EN (Chmiel et al. 2021)
- ESIC – EN, CZ, DE (Macháček et al. 2021)
- Further project in Belgrade, Louvain, Lisbon..



- **EPIC UdS** – EN, DE, ES

Transcription

Feature	Transcription
Silent pause	/
Filled pause	euh, hm, hum
Mid-word pause	spea/ euh ker [speaker]
Non-verbalized noise	[noise], [breath]
Non-standard pronunciation	report [repo:rt]
Inaudible segment	[inaudible]
Mispronunciation	plemary [plenary]
Truncated word	propo/
Ambiguity	they [?there]
Sentence-equivalent unit	↵

Transcription for interactional and non-verbal acoustic features based on EPICG

Metadata

Metadata	
Speaker related	Name Gender Nationality
Interpreter related	Gender Native language
Speech event related	General topic Title of debate Date Speech length in words Speech length in seconds Speech delivery rate (wpm) Speech delivery rate: *slow* ≤130w/m; *medium* = 131-160w/m; *high* ≥161w/m Source delivery type: read, impromptu, mixed

EPIC UdS – corpus variants

	no of tokens	no of sentences
ORG EN	67,526	3,622
SI EN DE	57,532	4,076
ORG DE	56,488	3,409
SI DE EN	58,503	3,623
ORG ES	53,947	2,537
SI ES EN	54,630	3,076

EPIC UdS – corpus variants

Comparable:

- EPIC UdS V2
  - incl. spoken features
- EPIC UdS V3
  - w/o some spoken features

	no of tokens	no of sentences
ORG EN	67,526	3,622
SI EN DE	57,532	4,076
ORG DE	56,488	3,409
SI DE EN	58,503	3,623
ORG ES	53,947	2,537
SI ES EN	54,630	3,076

Parallel:

- Sentence aligned
  - aligned, POS (incl. spoken features)
  - aligned, parsed (w/o some spoken features)
  - aligned, parsed surprisal (w/o some spoken features)

EPIC UdS: dependency parsing

EPIC UdS V2 (spacy v2.3.4)

EN: Stated accuracy evaluation: LAS = 90.28; UAS = 92.09

EPIC UdS EN V2 accuracy: **LAS = 78.73; UAS = 86.42**

DE: Stated accuracy evaluation: LAS = 91.15; UAS = 92.99

EPIC UdS DE V2 accuracy: **LAS = 69.74; UAS = 76.64**

EPIC UdS V3 (Stanza v1.2.3)

EN: EPIC EN V3 accuracy: **LAS = 85.94 ; UAS = 89.08**

DE: EPIC DE V3 accuracy: **LAS = 85.78; UAS = 91.03**

Labeled attachment score (LAS) = % of words with correct head and label

Unlabeled attachment score (UAS) = % of words with correct head

Applications: Comparable studies

Kullback-Leibler Divergence (KLD) on token-based n-gram models:

Interpreting

Original spoken

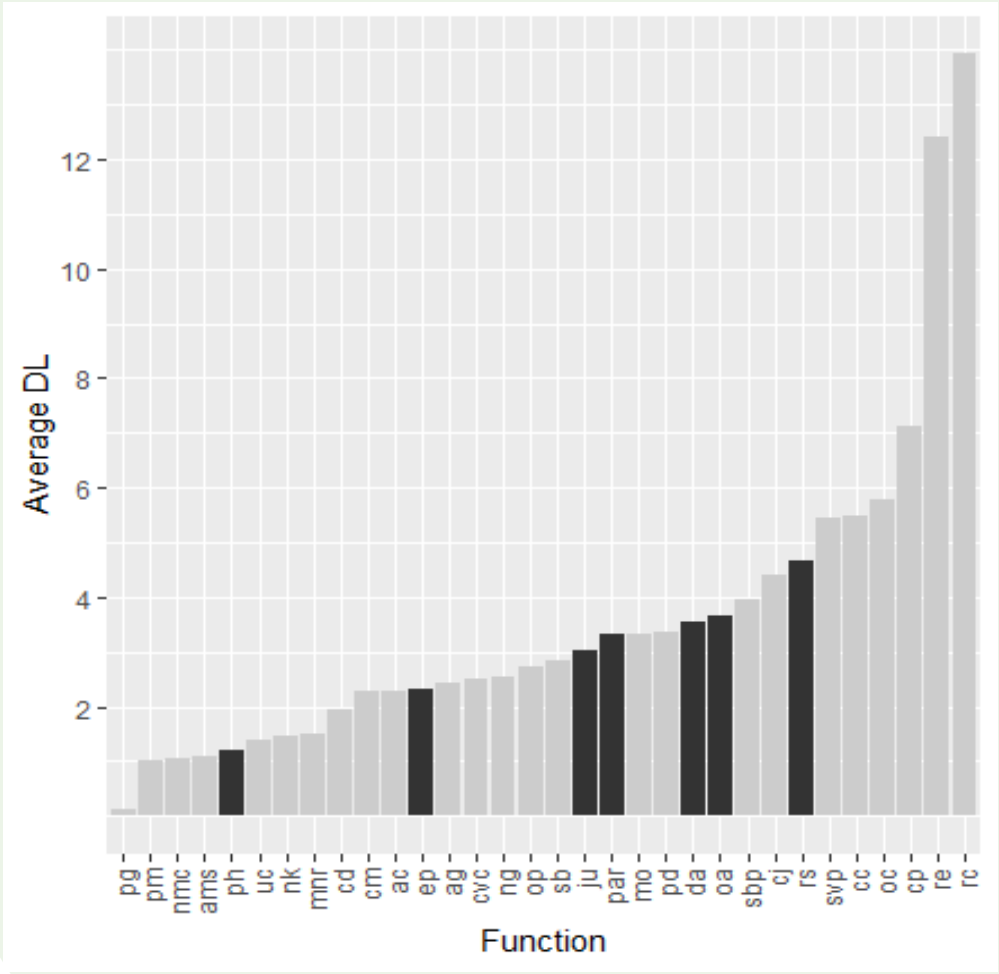
=> more high frequent and distinctive words

=> more less frequent and diverse words

(Przybyl et al. 2022)

Applications: Comparable studies

Syntactic complexity via dependency length:



- Syntactic complexity indicators:

  overall ORG more complex than SI

- especially longer relations are more frequent in ORG

: more frequent in ORG

: more frequent in SI

Applications: Parallel studies

Bilingual word embeddings with skipgram Word2Vec:

	Translation	Interpreting
Vocabulary size	18592	10524
Avg. Similarity	0.268	0.213
1st neighour	0.851	0.860
10th neighbour	0.723	0.685

Larger vocabulary size in TR

Comparable avg. similarity in both spaces (slightly higher in TR)

Comparable first neighbour similarity in both space (slightly higher in SI)

Higher 10th neighbour similarity in TR: SI clusters get looser faster

(Bizzoni and Teich 2019)

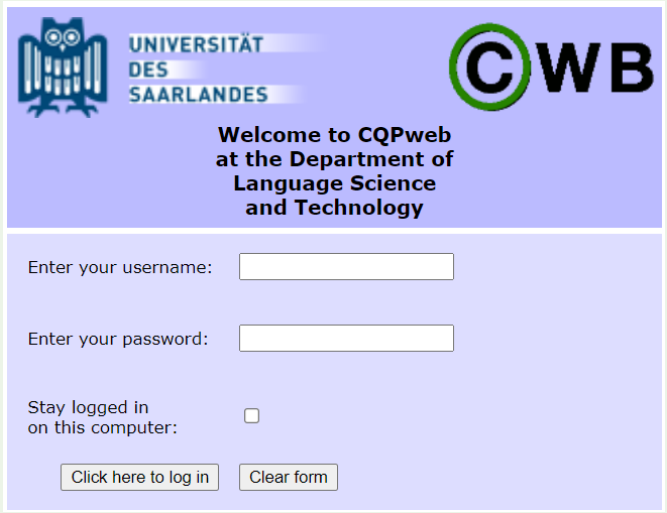
Available at

- Download of EPIC UdS V2:

http://hdl.handle.net/21.11119/0000-0008-F519-8

- Accessible online:

https://corpora.clarin-d.uni-saarland.de/cqpweb/



Thank you!

- heike.przybyl@uni-saarland.de
- e.lapshinova@mx.uni-saarland.de
- k.menzel@mx.uni-saarland.de
- stefan.fischer@uni-saarland.de
- e.teich@mx.uni-saarland.de