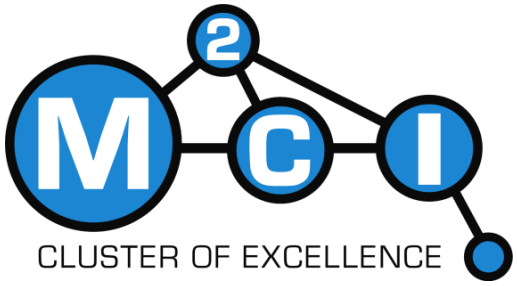


Universität
des Saarlandes

Universität
Heidelberg

Linking discourse modes and situation entity types in a cross-linguistic corpus study

Kleio-Isidora Mavridou, Annemarie Friedrich,
Melissa Peate Sørensen, Alexis Palmer and Manfred Pinkal

LSDSem Workshop, Lisbon, 2015

Discourse modes and situation entity types

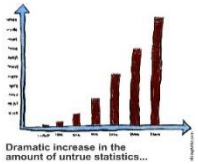
[Smith, 2003]

DISCOURSE MODES (per text passage)

NARRATIVE

DESCRIPTION

INFORMATION



REPORT

ARGUMENT/ COMMENTARY

- ≠ genre
- distinct linguistic characteristics

SITUATION ENTITIES (per clause)

STATE John is tall.

EVENT He hit his head on the door.

GENERALIZING SENTENCE He often does that.

GENERIC SENTENCE Tall people are clumsy.

QUESTION Have you seen John?

IMPERATIVE Look at him!

Poster!

QUESTIONS

Cross-linguistic correspondence of discourse modes (DMs)

Cross-linguistic correspondence of situation entity (SE) types

Distribution of SE types per DM

[Smith (2003), Palmer et al. (2007), Friedrich and Palmer (2014)]

Overview of this work

parallel corpus

Annotation study:
discourse modes

Corpus study:
situation entities

intuitively annotate passages with DMs

correspondence between DMs and SE types

adjust EN annotation scheme [Friedrich and Palmer (2014)] to DE

WHY?

empirical analysis for linguistic theory of DMs

improving NLP applications

translation studies
machine translation

Corpus data

11 parallel texts

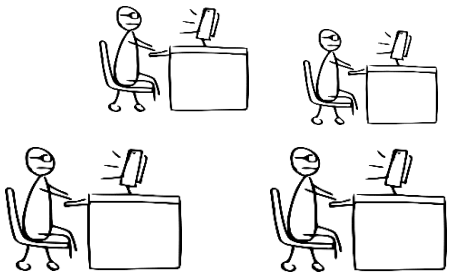

NewsCommentary + Global Voices
OPUS Books [Tiedemann, 2012]
Europarl [Islam and Mehler, 2012]
Smultron "Economy texts" and "Sophie's World" [Volk et al., 2010]

Corpus section	# tokens	# clauses	# paragraphs
Smultron aligned	10191 (en) 10719 (de)	1028	372
Other aligned	7115 (en) 6890 (de)	761	118
Total aligned	**17306 (en)** **17609 (de)**	**1789**	**490**

segmentation into paragraphs & clauses

SPADE [Soricut and Marcu, 2003] for EN,
syntax-based discourse segmenter for DE

annotation



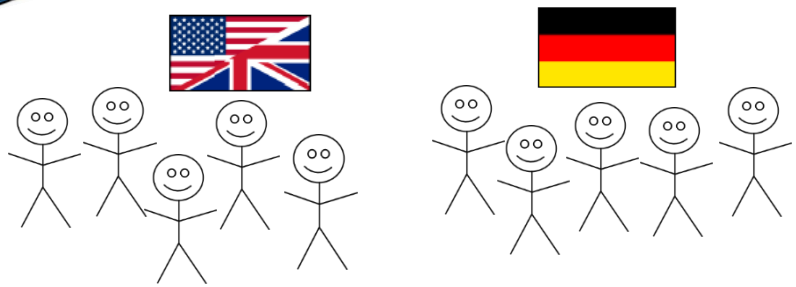
alignment

Paragraphs
manual

clauses
Smultron: semi-automatic
Other: manual

Pilot discourse modes annotation study

First corpus with paragraphs labeled with discourse modes!

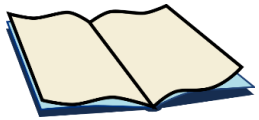


Fleiss' κ

English	German
0.46	0.50

Agreement chunk:
50 paragraphs

Short manual with intuitive descriptions
(1 prototypical example per mode)
+ short training



Annotation:
3-7 chunks of 30 paragraphs per annotator (490 paragraphs in total)

DIFFICULTIES

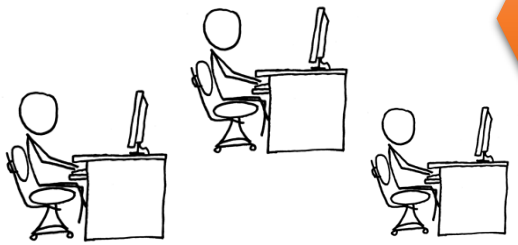
- linguistic characteristics not specified in manual
- distinction between DMs and genre
- DM boundaries vs. paragraph boundaries

Basis for studying discourse modes empirically

Situation entities corpus study

Adjust existing English annotation scheme of Friedrich and Palmer (2014) to German

Poster!



extensive manual + training

majority voting

gold standard

Fleiss' κ

Corpus section	English	German
Smultron	0.63	0.62
Other	0.61	0.67

DIFFERENCE: past/present perfect

English perfect = stative [Katz, 2003]
I have eaten. **(STATE)**

German:

(a) Ich habe schon gegessen. **(STATE)**
I have eaten.

(b) Gestern sind wir ins Kino gegangen. **(EVENT)**
Yesterday we went to the movies.

(c) Sie haben mir den Job gegeben.
They gave me the job. / They have given me the job.

⇨ **EVENT-PERFECT-STATE**

How consistently do German native speakers interpret clauses containing **perfect**?
⇨ **LARGE SCALE ANNOTATION EXPERIMENT**

Poster!

Cross-linguistic correspondence of SE types

		German							
		STATE	EVENT	EVT-PERF-ST	GENERAL.	GENERIC	IMP.	QUEST.	-
English	STATE	642	85	27	14	47	0	4	34
	EVENT	40	304	14	10	5	1	0	9
	GENERAL.	9	5	0	38	49	1	0	6
	GENERIC	33	0	0	1	143	0	0	3
	IMP.	2	1	0	0	0	9	0	2
	QUESTION	2	0	0	0	1	0	62	5
	-	57	32	2	8	41	0	4	37

QUALITATIVE ANALYSIS

40% of mismatches: **"general noise"**

identifying generics → hard also in monolingual setting [Friedrich et al., 2015]

60% of mismatches: **language-pair specific**

She **was startled**. (STATE)
Sie **fuhr zusammen**. (EVENT)

Take a look at… (IMPERATIVE)
Hier **können** Sie … **sehen**. (STATE)

Poster!

SE types correspond cross-linguistically, many shifts are systematic

Distributions of SE types per DM

PREDICTIONS [Smith, 2003]

NARRATIVE →
EVENTS and STATES

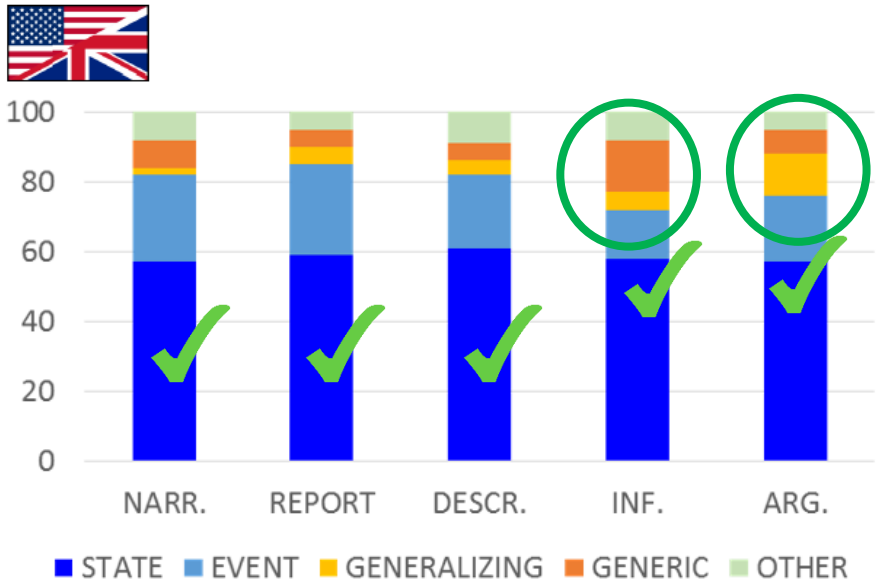
REPORT →
EVENTS, STATES,
GENERALIZING and GENERIC

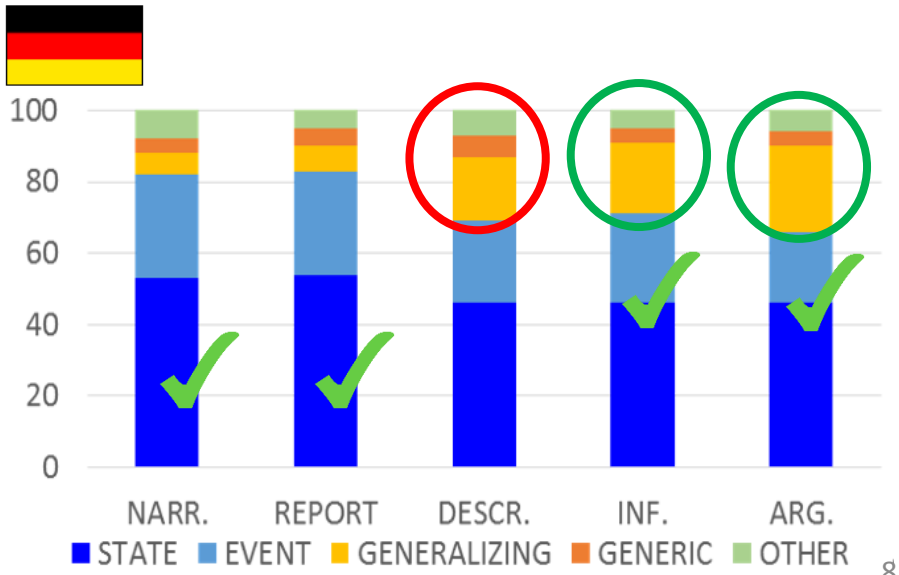
DESCRIPTION →
EVENTS and STATES

annotation problem?

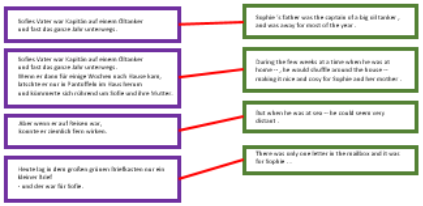
INFORMATION and **ARGUMENT →**
GENERALIZING and GENERIC

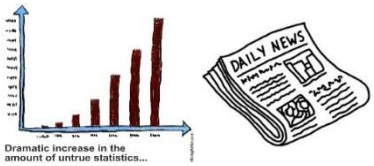
SE distribution similar across languages for most DMs



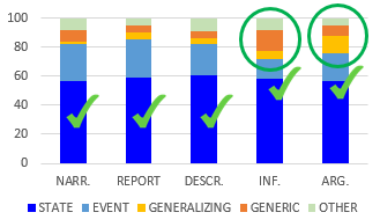


Conclusion and future work





English	German							
	STATE	EVENT	EVT-PERF-ST	GENERAL.	GENERIC	IMP.	QUEST.	-
STATE	642	85	27	14	47	0	4	34
EVENT	40	304	14	10	5	1	0	9
GENERAL.	9	5	0	38	49	1	0	6
GENERIC	33	0	0	1	143	0	0	3
IMP.	2	1	0	0	0	9	0	2
QUESTION	2	0	0	0	1	0	62	5
-	57	32	2	8	41	0	4	37



- ✓ First parallel corpus annotated for SEs, first corpus labeled with DMs
- ✓ Pilot DM annotation study – foundation for ongoing work
- ✓ SE types mainly correspond cross-linguistically and most shifts are systematic
- ✓ Labeled DMs mostly have the SE type distributions predicted by Smith (2003)

FUTURE WORK

- ★ **Computational models for SEs and DMs**
- ★ **Relevance for machine translation (evaluation)?**
- ★ **Analysis of additional languages**
- ★ **This work: translated texts -- distributions in original texts?**

Thank you!

Many thanks to:
Christine Bocionek, Fernando Ardente, Wladimir Sidorenko and to our many volunteer annotators for their time and help!

More info:
www.coli.uni-saarland.de/projects/sitent

References

Annemarie Friedrich and Alexis Palmer. 2014. **Situation entity annotation**. LAW VIII, page 149.

Zahurul Islam and Alexander Mehler. 2012. **Customization of the Europarl Corpus for Translation Studies**. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Graham Katz. 2003. **On the stativity of the English perfect**. *Perfect explorations*, pages 205-234.

Carlota Smith. 2003. **Modes of discourse: The local structure of texts**. Cambridge University Press.

Radu Soricut and Daniel Marcu. 2003. **Sentence level discourse parsing using syntactic and lexical information**. In *Proceedings of the 2003 Conference of NAACL-HLT.* Association for Computational Linguistics.

Jörg Tiedemann. 2012. **Parallel Data, Tools and Interfaces in OPUS**. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. 2010. **SMULTRON (version 3.0) —The Stockholm MULtilingual parallel TReebank**.