
DiscoGeM 2.0: A Parallel Corpus of English, German, French and Czech Implicit Discourse Relations

Frances Yung, Merel Scholman, Sarka Zikanova, Vera Demberg

Saarland University, Utrecht University, Charles University



Introduction

- Discourse relations (DRs) : semantic links between texts
- Can be **explicit** (marked with connectives) or **implicit** (unmarked)

Example:

1. I'm a feminist **because** I believe in gender equality.
2. I'm a feminist; **in other words**, I believe in gender equality.
3. I'm a feminist. I believe in gender equality.

- DR recognition is important for downstream NLP tasks, e.g. summarization.
- **Implicit** DR classification remains a challenge.
E.g. SOTA 14-way classification F1: 60% (GOLF, Jiang 2023)

Introduction

Challenges:

1. Lack of multi-lingual data
 - existing TED-MDB (Zeyrek 2019) only 200 implicit relations per language.
2. Lack of multi-domain data
3. DRs are highly ambiguous: soft label annotation preferred



DiscoGeM 2.0: A Parallel Corpus of GEnre-Mixed Implicit Discourse Relations

- 4 languages: **English, German, French, Czech**
- Parallel: **original vs translated** texts
- 2 domains: **Europarl & Literature**
- **Soft labels** by crowdsourcing: 10 PDTB3.0-labels per instance

DiscoGeM 1.0 (Scholman et al, 2022) vs DiscoGeM 2.0

- A corpus of genre-mixed implicit discourse relation in **English**

	DiscoGeM 1.0	DiscoGeM 2.0		
		Literature		
orig. ↓ / data lang. →	EN	DE	FR	CS
English (EN)	800	787	758	777
German (DE)	800	683	—	—
French (FR)	780	—	729	—
Czech (CS)	680	—	—	526
		Europarl		
English (EN)	418	417	414	—
German (DE)	701	701	—	—
French (FR)	739	—	727	—
Czech (CS)	700	—	—	697
Total parallel	5618	2588	2628	2000
		Wikipedia		
English (EN)	645	—	—	—

sentence alignment: Vec-align + LASER (Thompson and Koehn, 2019; Artetxe and Schwenk, 2019)

DiscoGeM 2.0

1. Methodology of the annotation
2. Annotation results

This talk

1. Methodology of the annotation
 - Background: DiscoGeM 1.0 (English)
 - Adaptation to other languages
2. Annotation results

This talk

1. Methodology of the annotation
 - Background: DiscoGeM 1.0 (English)
 - Adaptation to other languages
2. Annotation results

Methodology: background

- DiscoGeM 1.0 was crowdsourced by the **Two-step Discourse Connective Method** (Yung et al 2019)

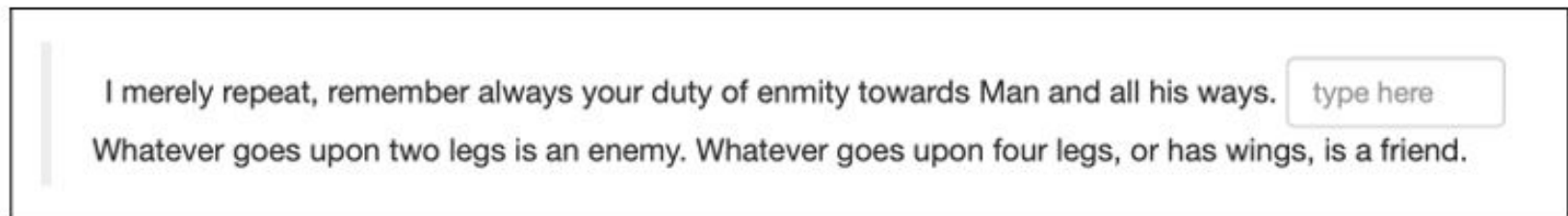
1. Freely insert a connective to express the relation

I merely repeat, remember always your duty of enmity towards Man and all his ways.
Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

Methodology: background

- DiscoGeM 1.0 was crowdsourced by the **Two-step Discourse Connective Method** (Yung et al 2019)

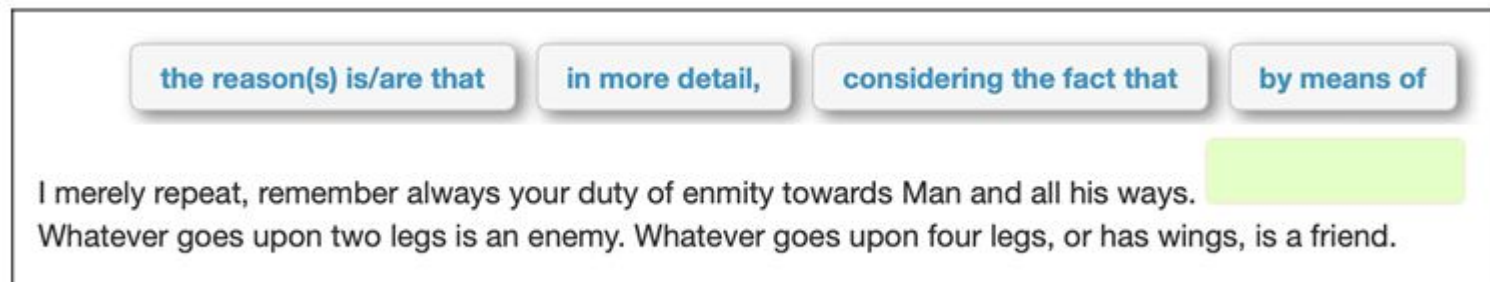
1. Freely insert a connective to express the relation



I merely repeat, remember always your duty of enmity towards Man and all his ways.

Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

2. Choose from a dynamic list to disambiguate



I merely repeat, remember always your duty of enmity towards Man and all his ways.

Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

This talk

1. Methodology of the annotation
 - Data: DiscoGeM 1.0 (English)
 - Adaptation to other languages
2. Annotation results

Methodology: motivation

- Insertion of connectives often requires a **change in word order** in other languages.

EN:

I'm feminist ...

because / in other words I believe in gender equality.

DE:

Ich bin Feministin ...

- **weil** ich an die Gleichstellung der Geschlechter glaube.
- **anders gesagt**, ich glaube an die Gleichstellung der Geschlechter.

- Crowdworkers may avoid connectives that lead to ungrammatical sequences irrespective of the meaning.

Methodology: One-step Connective Insertion

- More emphasis on the **semantic relation** expressed by the connective than whether it “fits” *syntactically* in context.

One day she left the same way. She came with a heavy suitcase. She left with a heavy suitcase. He paid the bill, left the restaurant and started walking through the streets, his melancholy growing more and more beautiful.

Drag and drop the word / phrase that links the highlighted texts:

Next



- The answer box is located outside the text.
- Specific note in the task instructions:

Focus on the meaning of linking words. You don't have to consider if it is grammatically correct or natural to insert that word between the highlighted texts.

Methodology: One-step Connective Insertion

- A **static list of connectives** to choose from, instead of free insertion.
- Each corresponds to one DR defined in PDTB 3.0.
- Semantically grouped for easier navigation.

One day she left the same way. She came with a heavy suitcase. She left with a heavy suitcase. He paid the bill, left the restaurant and started walking through the streets, his melancholy growing more and more beautiful.

Drag and drop the word / phrase that links the highlighted texts:

Next

then

at the same time

after

for that purpose

because

unless

so that

in that case

as a result

if not

if

for example

in more detail

in short

also

this illustrates that

in other words

or

similarly

as if

rather than

instead

other than that

an exception is that

(no direct relation)

thereby

on the other hand

even though

nonetheless

Methodology: Multi-lingual connective list

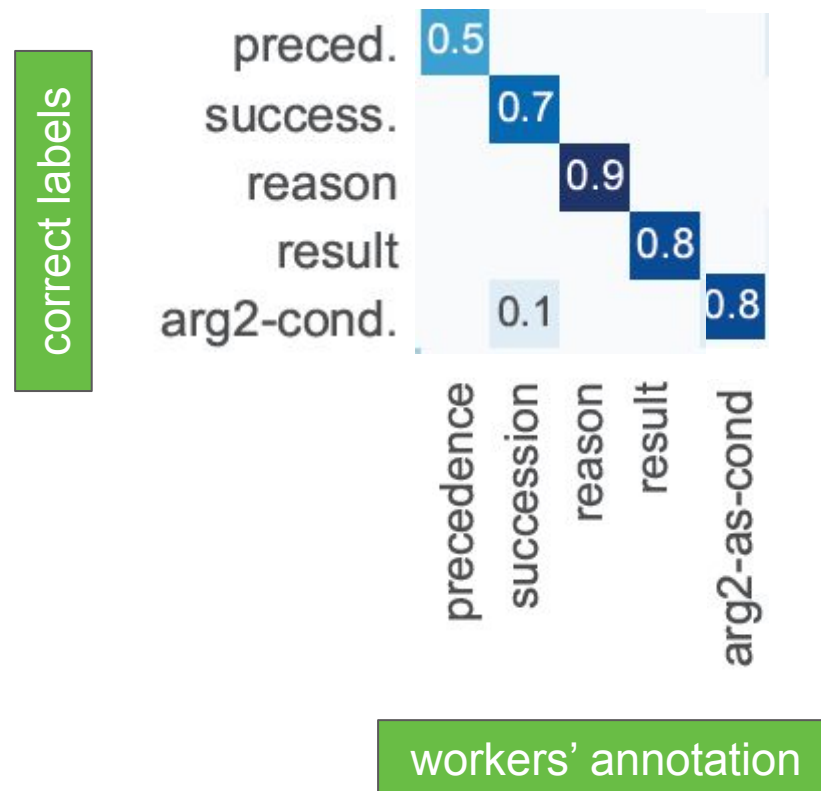
- A balance of
 - ambiguity - preference of single-sense connectives
 - frequency - avoidance of rare connectives
 - generalizability - avoidance of syntactic/stylistic dependent connectives
- Based on connective lexicons and consultation with native linguists.

Relation sense		English	German	French	Czech
TEMPORAL	PRECEDENCE	then	dann	ensuite	potom
	SUCCESSION	after	davor,	après que	předtím
	SYNCHRONOUS	at the same time	gleichzeitig	en même temps	zároveň
CAUSAL	REASON	because	weil	parce que	protože
	RESULT	therefore	daher	c'est pourquoi	proto
COMPARISON	ARG2-AS-DENIER	nonetheless	trotzdem	néanmoins	přesto
	CONTRAST	on the other hand	andererseits	d'autre part	na druhou stranu
EXPANSION	CONJUNCTION	also	darüberhinaus	en plus	také
	ARG2-AS-INST.	for example	zum Beispiel	par exemple	například
	ARG2-AS-DETAIL	in more detail	genauer gesagt	plus précisément	konkrétně
NO RELATION		(no direct relation)	(keine direkte Beziehung)	(pas de relation directe)	(bez přímého vztahu)

- See paper appendix for the full list.

Methodology: validation

- Procedure:
 - Native speakers of DE, FR, CS recruited on Prolific.
 - Screened by a **selection task** of 18 questions (Pass: $\geq 50\%$)

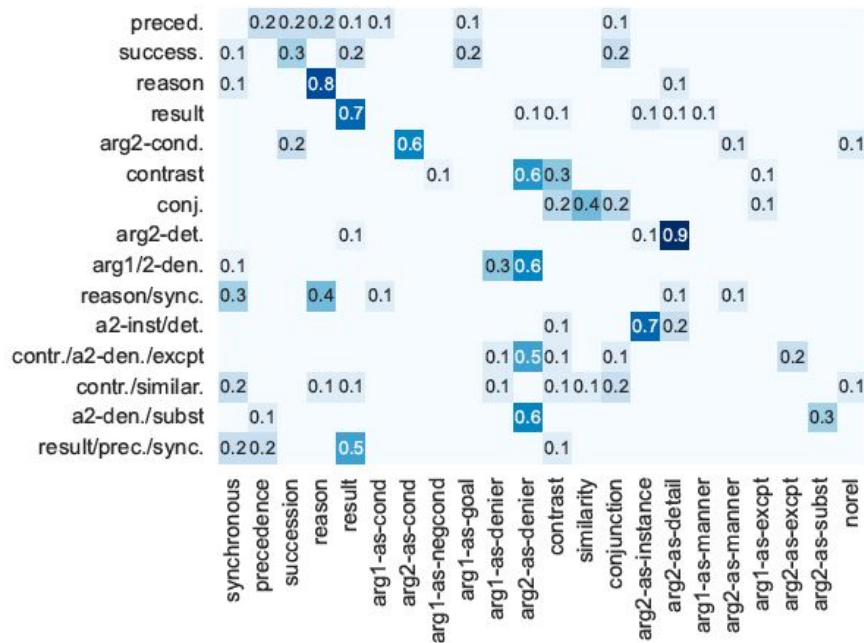


DE results

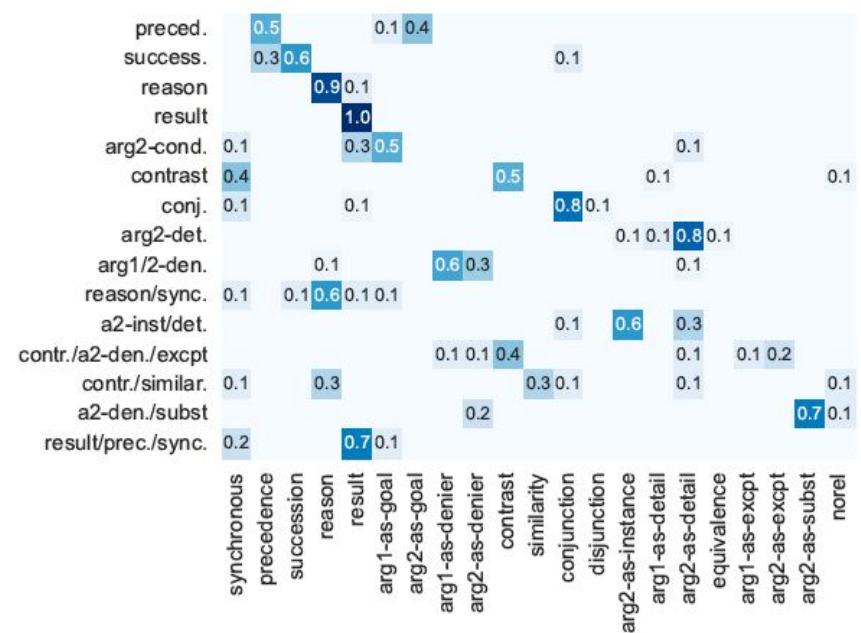
- High agreement for single-sense instances
- Multi-sense instances annotated by distributional labels.

Methodology: validation

- High agreement in the selection task in other languages as well, but depends on relations.
- Near perfect agreement between the two-step and one-step approach in English.
- Cross-lingual divergence in agreement; lexical gaps between connectives in different languages.



FR



CS

This talk

1. Methodology of the annotation
 - Data: DiscoGeM 1.0 (English)
 - Adaptation to other languages
2. Annotation results
 - General statistics
 - Cross-lingual comparison

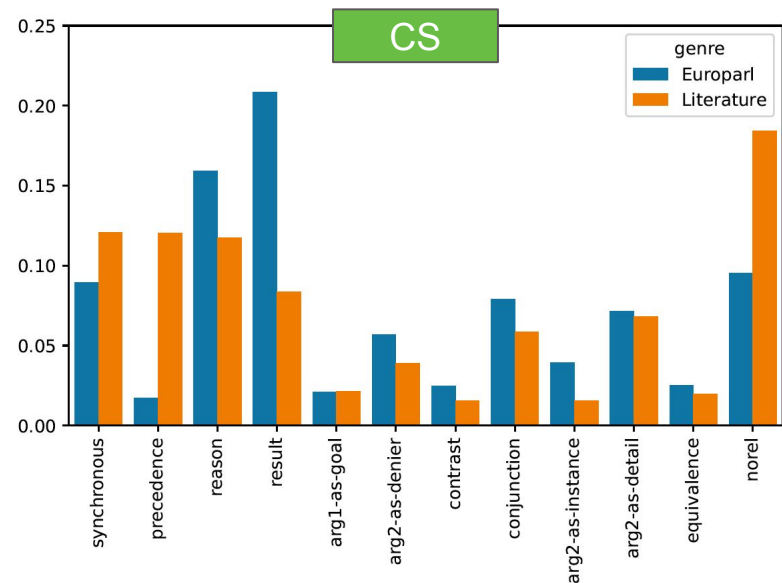
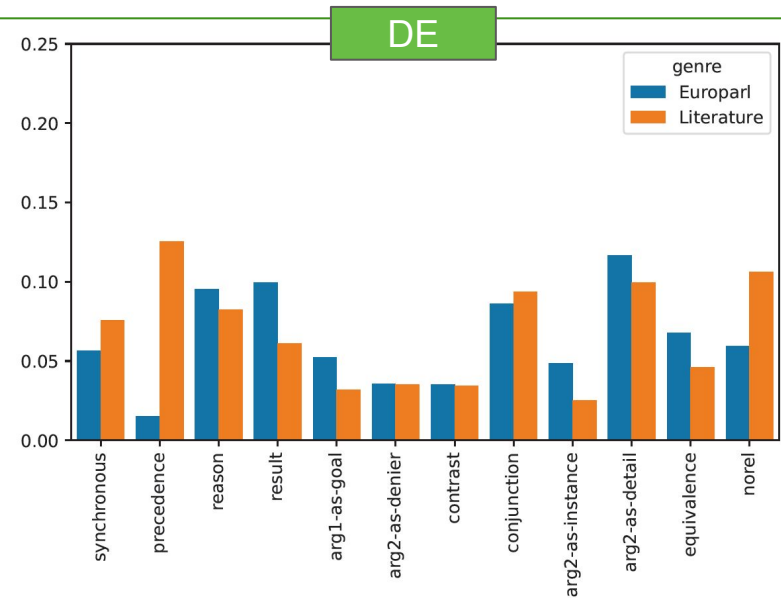
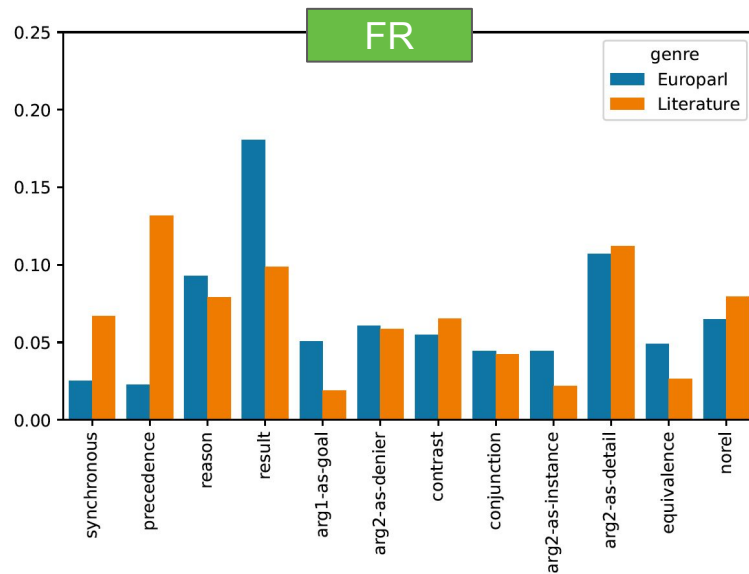
Results: statistics

	Europarl			
orig. ↓ / data lang. →	EN	DE	FR	CS
English (EN)	418	417	414	—
German (DE)	701	701	—	—
French (FR)	739	—	727	—
Czech (CS)	700	—	—	697
Subtotal	2558	1118	1141	697
	Literature			
orig. ↓ / data lang. →	EN	DE	FR	CS
English (EN)	800	787	758	777
German (DE)	800	683	—	—
French (FR)	780	—	729	—
Czech (CS)	680	—	—	526
Subtotal	3060	1470	1487	1303
Total	5618	2588	2628	2000

- 5,618 English items in DiscoGeM 1.0 → 12,834 multilingual items in DiscoGeM 2.0
- Translation to/from English
- Not all items were alignable (e.g. 2 EN sents translated to 1 DE sent)

Results: relation distribution in language subsets

- Genre effects observed in EN in DiscoGeM 1.0 appear also in other languages (DiscoGeM 2.0):
 - more “RESULT” in Europarl
 - more “PRECEDENCE” in literature



Results: majority labels of aligned relations

- General cross-lingual agreement
- Expected patterns of co-occurrence and confusion (e.g. “cause” & “level-of-details”; “concession” and “contrast”)
- Language specific patterns (e.g. fewer “cause” in DE)

Czech	synchronous	20	15	7	6	10	1	17	5
	asynchronous	5	91	16	1	3		7	
	cause	3	15	67	15	5	4	21	4
	concession	5	3	11	28			8	2
	contrast				5		1		
	instantiation			2	1	1	4		2
	level-of-detail	4	5	12	9	5	5	51	
	conjunction	1	10	4	1	9	2	10	9
		synchronous	asynchronous	cause	concession	contrast	instantiation	level-of-detail	conjunction

French (742 items)

French	synchronous	17	7	4	3			1	2
	asynchronous	5	113	7	3			4	14
	cause	4	21	122	13	3	3	37	17
	concession	5	5	15	38	13	2	16	15
	contrast	4	4	8	3	4	1	5	14
	instantiation			4	3		13	7	2
	level-of-detail	8	9	23	15	1	5	98	25
	conjunction	2		10			1	1	14
		synchronous	asynchronous	cause	concession	contrast	instantiation	level-of-detail	conjunction

German (1162 items)

Czech	synchronous	18	19	7	6	2	1	6	16
	asynchronous	3	100	9		1		3	9
	cause	5	20	61	6	1	2	22	11
	concession	3	4	4	24	3		8	3
	contrast	1			2	2	1		
	instantiation		1	2	1		4		1
	level-of-detail	4	6	8	6		3	46	9
	conjunction	3	7		1	1	3	5	31
		synchronous	asynchronous	cause	concession	contrast	instantiation	level-of-detail	conjunction

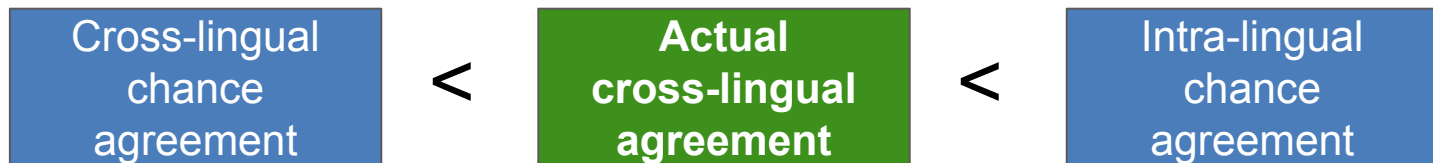
German (768 items)

Results: comparing the label distributions

- Evaluate by **Jensen-Shannon Divergence (JSD)** between the label distributions of the **same item** but **different languages**
 - a. **cross-lingual chance agreement**
JSD between unaligned and shuffled cross-lingual annotations
 - b. **intra-lingual chance agreement**
JSD between two sampled label distributions of a particular item
 - c. **actual cross-lingual agreement**
Actual JSD between the two language versions of the same item

Results: comparing the label distributions

- Evaluate by **Jensen-Shannon Divergence (JSD)** between the label distributions of the **same item** but **different languages**
 - a. **cross-lingual chance agreement** = 0.83 on average
JSD between unaligned and shuffled cross-lingual annotations
 - b. **intra-lingual chance agreement** = 0.43 on average
JSD between two sampled label distributions of a particular item
 - c. **actual cross-lingual agreement** = 0.63~0.71 on average
Actual JSD between the two language versions of the same item



(Lower JSD = higher agreement)

Results: examples

EXAMPLE 1:

Original German text: Du sollst aber nie vergessen, was ich dir so oft gesagt habe: unsere Bestimmung ist, die Gegensätze richtig zu erkennen, erstens nämlich als Gegensätze, dann aber als die Pole einer Einheit. // So ist es auch mit dem Glasperlenspiel.

Translation by Deep Translate: *But you should never forget what I have told you so often : our destiny is to recognize the contrasts correctly, first of all as contrasts, but then as the poles of a unity. // So it is with the Glass Bead Game.*

Translated English text: But never forget what I have told you so often: our mission is to recognize contraries for what they are: first of all as contraries, but the opposite poles of a unity. // Such is the nature of the Glass Bead Game.

- **Annotated labels on German:**

SIMILARITY (5), REASON (2), EQUIVALENCE (2), CONTRAST (1)

- **Annotated labels on English:**

ARG1-AS-DETAIL (6), RESULT (3), CONJUNCTION (1)

EXAMPLE 2:

Original German text: Ich hatte sie noch nie mit Hut gesehen, sie hatte sich immer geweigert, einen aufzusetzen. Der Hut veränderte sie sehr. // Sie sah wie eine junge Frau aus. Ich dachte, sie mache einen Ausflug, obwohl es eine merkwürdige Zeit für Ausflüge war.

Translation by Deep Translate: *I had never seen her in a hat before, she had always refused to wear one. The hat changed her a lot. // She looked like a young woman. I thought she was going on an outing, although it was a strange time for outings.*

English translated text: I had never seen her in a hat before, she had always refused to wear one. The hat altered her very much. // She looked like a young woman. I thought she must be going on an outing, though it was a strange time for outings. But in those days the schools were capable of anything.

- **Annotated labels on German:**

REASON (5), EQUIVALENCE (2), ARG1-AS-DETAIL (1), ARG2-AS-GOAL (1), NO RELATION (1)

- **Annotated labels on English:**

RESULT (7), CONTRAST (1), ARG2-AS-INSTANCE (1), PRECEDENCE (1)

Conclusion

- A discourse-annotated corpus unlike any others.
- Download: <https://github.com/merelscholman/DiscoGeM>
- Cross-lingual comparison reveals that implicit DR annotations are not always projectable.
- Further analysis is required to investigate the reasons behind the cross-lingual disagreement.

Thank you

The research reported in this paper was supported by the German Research Foundation (DFG) under Grant SFB 1102 (“Information Density and Linguistic Encoding”, Project- ID 232722074) and the Czech Science Foundation (project no. 24-11132S, Disagreement in Corpus Annotation and Variation in Human Understanding of Text); a part of the used data comes from the project no. LM2018101 by the Czech Ministry of Education, Youth and Sports (Digital Research Infrastructure for Language Technologies, Arts and Humanities)