UNIVERSITÄT DES SAARLANDES

Analysis and Annotation of Coreference for Contrastive Linguistics and Translation Studies

Ekaterina Lapshinova-Koltunski

IPI PAN, Warszawa

December 3, 2018

Overview

UNIVERSITÄT DES SAARLANDES

Coreference

UNIVERSITÄT DES SAARLANDES

Corefential/coreference chain

all the mentions of one and the same entity (also abstract: event, state) through the whole text

Police say a husband fatally shot his wife and another man before killing himself in a central Pennsylvania motel room. The York County Coroner's Office says 35-year-old Donnell Graham shot his wife.

Entity: Donnell Graham:

- indefinite noun *a husband* (antecedent)
- possessive pronoun *his*
- reflexive pronoun *himself*
- NP including a proper name *35-year-old Donnell Graham*
- some frameworks: a zero inexpressible pronoun as a subject of a clause with a gerund *killing*

Our Interest: English and German

UNIVERSITÄT DES SAARLANDES

- full coreference chains (NPs, pronouns, other means)
- further discourse-related phenomena (e.g. DRDs)

- **Contrastive Linguistics:** differences in the range of linguistic means triggering the relation
  [Kunz and Steiner, 2012a, Kunz and Lapshinova-Koltunski, 2015]
- **Translation Studies**: differences in their realisation $\Rightarrow$ transformation patterns: references in the source language (SL) must be rendered with appropriate linguistic devices from the repertoire of the target language (TL), with different constraints
- **Resource Compilation:** annotated for English-German:
  GECCo [Lapshinova-Koltunski and Kunz, 2014],
  ParCor [Guillou et al., 2014]
  corpus by [Grishina and Stede, 2015]

Empirical Analysis of Discourse Phenomena

UNIVERSITÄT DES SAARLANDES

1. Top-down / Theory-driven

Existing theoretical knowledge (e.g. from grammars) is used as a background for the analysis (annotation) of discourse phenomena

2. Bottom-up / Data-driven

Start from the data and use the existing theories / frameworks to interpret the findings

1. Top-down

UNIVERSITÄT DES SAARLANDES

- existing theories / grammars are used for the formulation of the categories existing in both languages
- these categories are operationalised in terms of linguistic features, i.e. lexico-grammatical patterns (coreference expressed in pronouns, nominal phrases, etc.)
- these categories are then annotated in parallel and comparable corpora
- existing resources are used for an empirical analysis (e.g. frequency-based patterns of use) in contrastive linguistics and translation or used as training data in NLP

$\Rightarrow$ **THEORY-DRIVEN STUDIES**

2. Bottom-up

UNIVERSITÄT DES SAARLANDES

- start with the data in parallel or comparable corpora
- use shallow / “easy-to-get” structures (token-, pos-, parser-output-based)
- extract patterns of the “easy-to-get” structures
- try to interpret them with the help of theoretical frameworks
- existing theories facilitate interpreting quantitative and qualitative results

$\Rightarrow$ **DATA-DRIVEN STUDIES**

UNIVERSITÄT DES SAARLANDES

Example Study I: GECCo

Research project supported by DFG (GZ STE 840/6-2 and KU 3129/1-2)

Web: www.gecco.uni-saarland.de

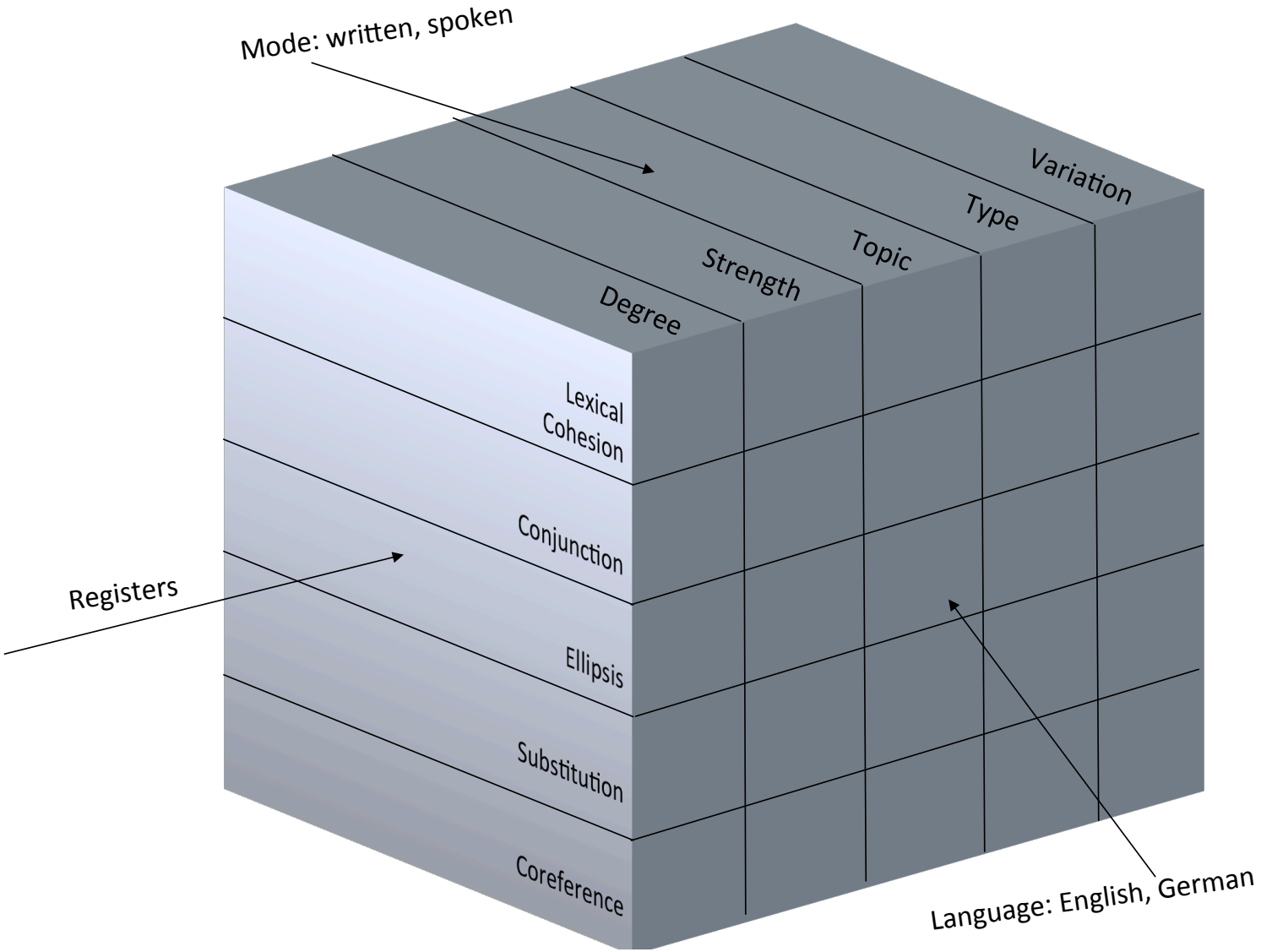
- Kerstin Kunz (Uni Heidelberg), Erich Steiner
- José M. Martínez Martínez, Stefania Degaetano-Ortlieb, Marilisa Amoia

Framework: Cohesion

UNIVERSITÄT DES SAARLANDES

Types of Cohesion (cf. Halliday & Hasan 1976)	**Meaning relations**
Coreference *An option … it/ this option*	identity
Substitution *Many options … a good one*	Type reference/ comparison
Ellipsis *You will feel disappointment. []* *Maybe.* *Many options … a good [].*	
Comparative Reference *One option … another/ better option*	
Cohesive conjunction *X. But/ And/ However Y*	Logico-semantic relations (addition, contrast, cause, …)
Lexical cohesion *factors … factors*	Similarity (repetition, general nouns, meronymy, Synonymy …)

Research Design



Theoretical Background

UNIVERSITÄT DES SAARLANDES

- Language: EO vs. GO
  (Hawkins, 1986; König&Gast, 2012; Königs, 2011, etc.)
- Mode of production: spoken vs. written
  (Biber, 1988; Mair, 2006; Leech et al., 2009)
- Register
  (Biber, 1995; House 2002; Hansen-Schirra et al., 2012; Neumann, 2013)

Empirical Methodology

UNIVERSITÄT DES SAARLANDES

- define operationalisations:
  relate cohesive features to abstract dimensions
- for each variable (language, register, mode):
  - extract instances/frequencies for cohesive features from a corpus
  - evaluate frequencies statistically
- interpret results in terms of abstract dimensions

Operationalisations

UNIVERSITÄT DES SAARLANDES

1. Degree of Cohesion:
   What is the average proportion of cohesive devices per text?

2. Strength of Cohesive Relations:
   - How explicit are cohesive devices?
   - How close are elements in cohesive chains?
   - How long are cohesive chains?

3. Types of Meaning Relations:
   - Which meaning relations are more important in languages/modes/registers?
   - Which meaning relations are most important for the distinction of languages/modes/registers

4. Breadth of variation: How much cohesive variation is there for
   - language: English <=> German
   - mode: spoken <=> written
   - register

Corpus Resources

UNIVERSITÄT DES SAARLANDES

GECCOCOH

register	EO	GO
ACADEMIC	40.559	43.703
ESSAY	34.998	35.668
FICTION	36.996	36.778
INSTR	36.167	36.880
INTERVIEW	37.898	40.198
POPSCI	35.148	36.177
SHARE	35.824	35.235
SPEECH	35.062	35.399
TOU	35.907	36.574
WEB	36.119	35.779
TOTAL	**364.678**	**372.391**

GECCOCHAIN

register	EO	GO
ESSAY	27.171	31.407
FICTION	36.996	36.778
INTERVIEW	30.057	35.036
POPSCI	27.055	32.639
TOTAL	**121.279**	**135.860**

- both subsets of GECCo
  (Lapshinova et al., 2012 and Hansen-Schirra et al., 2012)
- **GECCo annotation levels**
  word: ⇒ *word, lemma, pos*, **chunk**:⇒ *sentences, syntactic chunks, clauses*,
  text: ⇒ *registers*, **extralinguistic**: ⇒ *register analysis, speaker information*,

  cohesion (Lapshinova & Kunz, 2014; Martinez Martinez et al., 2016)

Annotation Procedures

UNIVERSITÄT DES SAARLANDES

[Lapshinova-Koltunski and Kunz, 2014]

- CWB perl modules
- based on YAC recursive chunker (Kermes and Evert, 2002; Kermes, 2003)
- automatic extraction and annotation of candidates
- manual correction

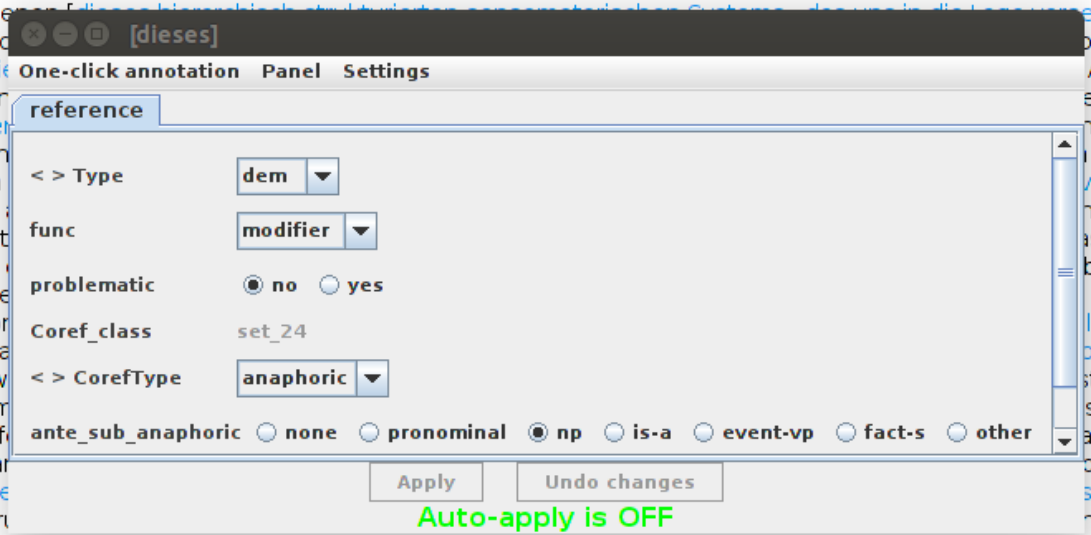
<reference type="dem" func="pronadv">
daraus
< /reference>

<reference type="dem" func="local" >
hier
< /reference>

<reference type="comp" func="particular">
grössere
< /reference>

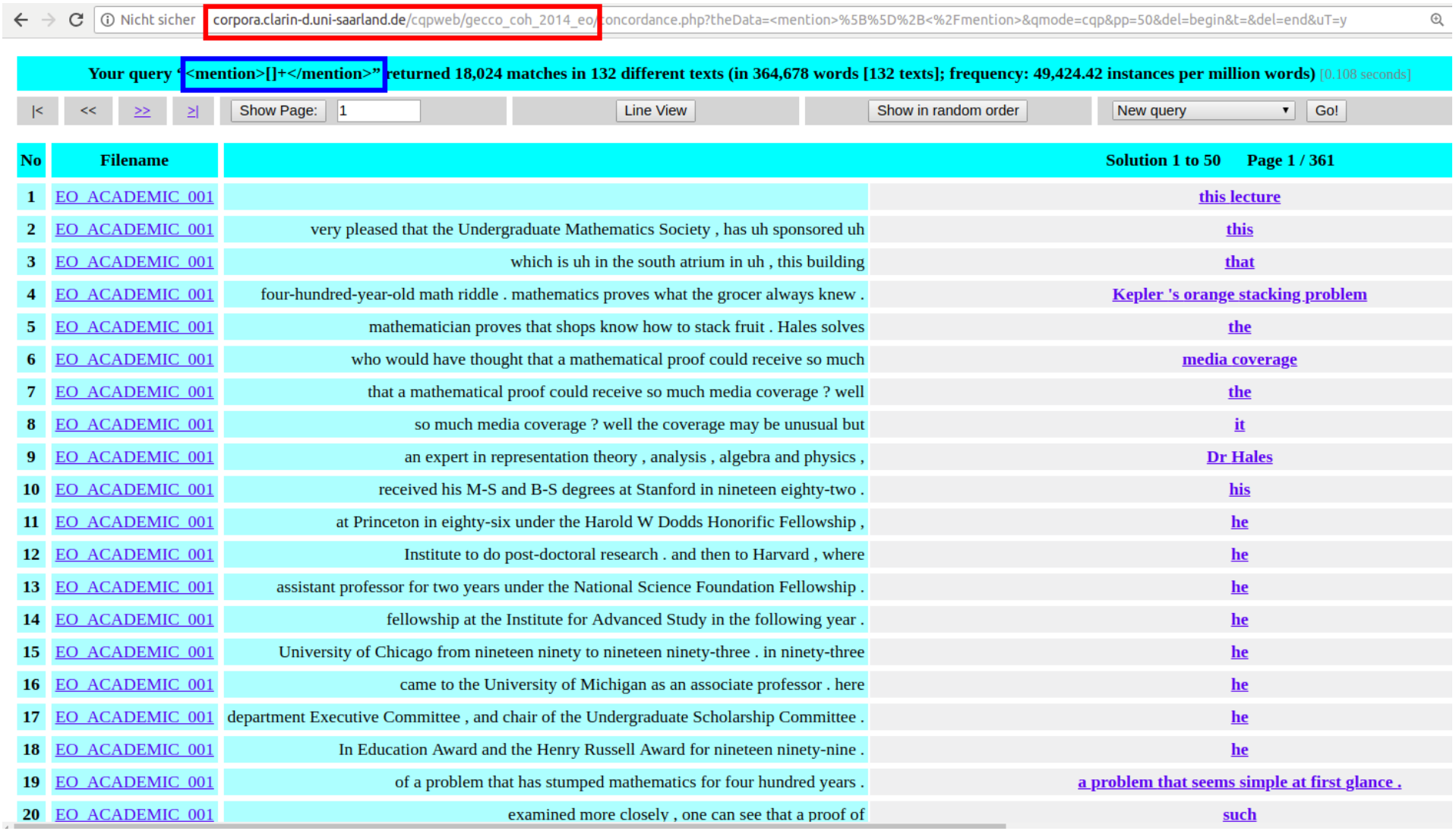
Annotation Procedures

- add antecedent & evaluate automatic system



(Müller & Strube, 2006)

Corpus Availability



Nicht sicher corpora.clarin-d.uni-saarland.de/cqpweb/gecco_coh_2014_eo/concordance.php?theData=<mention>%5B%5D%2B<%2Fmention>&qmode=cqp&pp=50&del=begin&t=&del=end&uT=y

Your query "<mention>[]+</mention>" returned 18,024 matches in 132 different texts (in 364,678 words [132 texts]; frequency: 49,424.42 instances per million words) [0.108 seconds]

|< << >> >| Show Page: 1 Line View Show in random order New query Go!

No	Filename		Solution 1 to 50 Page 1 / 361
1	EO_ACADEMIC_001		this lecture
2	EO_ACADEMIC_001	very pleased that the Undergraduate Mathematics Society , has uh sponsored uh	this
3	EO_ACADEMIC_001	which is uh in the south atrium in uh , this building	that
4	EO_ACADEMIC_001	four-hundred-year-old math riddle . mathematics proves what the grocer always knew .	Kepler 's orange stacking problem
5	EO_ACADEMIC_001	mathematician proves that shops know how to stack fruit . Hales solves	the
6	EO_ACADEMIC_001	who would have thought that a mathematical proof could receive so much	media coverage
7	EO_ACADEMIC_001	that a mathematical proof could receive so much media coverage ? well	the
8	EO_ACADEMIC_001	so much media coverage ? well the coverage may be unusual but	it
9	EO_ACADEMIC_001	an expert in representation theory , analysis , algebra and physics ,	Dr Hales
10	EO_ACADEMIC_001	received his M-S and B-S degrees at Stanford in nineteen eighty-two .	his
11	EO_ACADEMIC_001	at Princeton in eighty-six under the Harold W Dodds Honorific Fellowship ,	he
12	EO_ACADEMIC_001	Institute to do post-doctoral research . and then to Harvard , where	he
13	EO_ACADEMIC_001	assistant professor for two years under the National Science Foundation Fellowship .	he
14	EO_ACADEMIC_001	fellowship at the Institute for Advanced Study in the following year .	he
15	EO_ACADEMIC_001	University of Chicago from nineteen ninety to nineteen ninety-three . in ninety-three	he
16	EO_ACADEMIC_001	came to the University of Michigan as an associate professor . here	he
17	EO_ACADEMIC_001	department Executive Committee , and chair of the Undergraduate Scholarship Committee .	he
18	EO_ACADEMIC_001	In Education Award and the Henry Russell Award for nineteen ninety-nine .	he
19	EO_ACADEMIC_001	of a problem that has stumped mathematics for four hundred years .	a problem that seems simple at first glance .
20	EO_ACADEMIC_001	examined more closely , one can see that a proof of	such

Linguistic Features

UNIVERSITÄT DES SAARLANDES

COREFERENCE	SUBSTITUION	CONJUNCTION	ELLIPSIS
all antecedents	subst-nom, subst-verb, subst-claus	conj-addit-conn, conj-adversat-conn, conj-causal-conn, conj-addit-subjun, conj-adversat-subjun, conj-causal-subjun, conj-temp-subjun, conj-addit-adverb, conj-adversat-adverb, conj-causal-adverb, conj-temp-adverb, conj-modal-adverb	elli-antecedents
antecedent-np, antecedent-pronominal, antecedent-fact-s, antecedent-event-vp, antecedent-is-a, antecedent-other		conj-addit, conj-adversat, conj-causal, conj-temp, conj-modal	all elli
all anaphors		conj-conn, conj-subjun, conj-adverb	elli-nom, elli-verb, elli-claus, elli-yn, elli-mix
anaphors-pers-it, anaphors-pers-head, anaphors-pers-mod, anaphors-dem-head, anaphors-dem-mod, anaphors-dem-artic,anaphors-dem-pronadv, anaphors-dem-local,anaphors-dem-temporal, anaphors-comp-general,anaphors-comp-particular			
antecedent-subj, antecedent-obj, anaphors-subj, anaphors-obj			

LEXICAL COHESION	CHAINS
meaning relations	number-of-chains, length, switch-rate, distance

Analyses and Results

UNIVERSITÄT DES SAARLANDES

- **Coreference:** Kunz, K. & E. Steiner (2012). Towards a comparison of cohesive reference in English and German: System and text. In: M. Taboada, et al. (eds.), Contrastive Discourse Analysis. Functional and Corpus Perspectives. London: Equinox.
- **Substitution:** Kunz, K. & E. Steiner (2013). Cohesive substitution in English and German: a contrastive and corpus-based perspective, in: Aijmer, K. & Altenberg, B. eds. 2013. Advances in corpus-based contrastive linguistics. Studies in honour of S. Johansson. Amsterdam: John Benjamins. pp. 201–231.
- **Conjunction:** Kunz, K. & E. Lapshinova-Koltunski (2014). Cohesive conjunctions in English and German: Systemic contrasts and textual differences. In: Vandelanotte, L. et al. (eds). Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora. Amsterdam/New York: Rodopi, pp. 229-262.
- **Ellipsis:** Menzel, K. (2016). Understanding English-German contrasts - a corpus-based comparative analysis of ellipses as cohesive devices, PhD dissertation, Saarland University.
- **Lexical Cohesion:** Kunz, K., E. Lapshinova-Koltunski, J. M. Martínez-Martínez, K. Menzel & E. Steiner (2018). Shallow features as indicators of English-German contrasts in lexical cohesion. In: Languages in Contrast.
- **Overall**: Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K. and Steiner, E. (2017). GECCo - an empirically-based comparison of English-German cohesion. In De Sutter, G. and Delaere, I. and Lefer, M.-A. (eds.). Empirical Translation Studies. New Theoretical and Methodological Traditions. TILSM series. Vol. 300. Mouton de Gruyter, pp. 265-312.

cf. also **http://www.gecco.uni-saarland.de/GECCo/deliverables.html**

Summary and Outlook

comparable corpus with full **annotation of coreference** (and further cohesive devices)
$\Rightarrow$ valuable resource with a variety of uses:

Applications so far:

- numerous contrastive studies within GECCo
- teaching

Future Applications

- help to study the **mechanisms involved in coreference translation** $\Rightarrow$ better understanding of the phenomenon
- a resource for **training and development** of **multilingual or monolingual coreference resolution** systems?

UNIVERSITÄT DES SAARLANDES

Example Study II: ParCorFull

EAMT sponsorship project

LREC-Paper: [Lapshinova et al., 2018]

Our Interest: English and German

UNIVERSITÄT DES SAARLANDES

- full coreference chains (NPs, pronouns, other means)
- further discourse-related phenomena (e.g. DRDs)

- **Contrastive Linguistics:** differences in the range of linguistic means triggering the relation
  [Kunz and Steiner, 2012a, Kunz and Lapshinova-Koltunski, 2015]
- **Translation Studies**: differences in their realisation $\Rightarrow$ transformation patterns: references in the source language (SL) must be rendered with appropriate linguistic devices from the repertoire of the target language (TL), with different constraints
- **Resource Compilation:** annotated for English-German:
  GECCo [Lapshinova-Koltunski and Kunz, 2014],
  ParCor [Guillou et al., 2014]
  corpus by [Grishina and Stede, 2015]

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Segmentation**: Annotated elements (markables): pronouns, nouns, NPs, elliptical constructions that are parts of a coreference pair (antecedent-anaphora), as well as verbal phrases or clauses being antecedents of event anaphora

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Segmentation**: Annotated elements (markables): pronouns, nouns, NPs, elliptical constructions that are parts of a coreference pair (antecedent-anaphora), as well as verbal phrases or clauses being antecedents of event anaphora
- **Types of antecedents – entities and events:**
  - Entities are represented by a pronoun or an NP
  - Events are represented by a VP, a clause or a set of clauses
  - Split antecedents (*prosperity and oportunity*)
  - no explicit antecedent: anaphora, but no specific antecedent

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Segmentation**: Annotated elements (markables): pronouns, nouns, NPs, elliptical constructions that are parts of a coreference pair (antecedent-anaphora), as well as verbal phrases or clauses being antecedents of event anaphora
- **Types of antecedents – entities and events:**
  - Entities are represented by a pronoun or an NP
  - Events are represented by a VP, a clause or a set of clauses
  - Split antecedents (*prosperity and oportunity*)
  - no explicit antecedent: anaphora, but no specific antecedent
- **Types of anaphora – pronouns and NPs:**
  - pronouns: personal, relative, reflexive and demonstrative including locations (*there, here*) and time (*then, now*) and pronominal adverbs (*gegen+das* $\rightarrow$ *dagegen*)
    Viele Amerikaner haben Probleme mit Rassismus; doch wir sind dagegen immun. (racism, against + this)

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Types of anaphora – pronouns and NPs:**
  - NPs: proper names (*Herr Almeida Freire*), nominal premodifiers, full NPs (with a definite article/demonstrative modifier) and NPs with quantifiers (*all people* in the meaning *all these people*)

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Types of anaphora – pronouns and NPs:**
  - NPs: proper names (*Herr Almeida Freire*), nominal premodifiers, full NPs (with a definite article/demonstrative modifier) and NPs with quantifiers (*all people* in the meaning *all these people*)
  - generic nouns co-refer with definite full NPs or pronouns, but not with other generics

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Types of anaphora – pronouns and NPs:**
  - NPs: proper names (*Herr Almeida Freire*), nominal premodifiers, full NPs (with a definite article/demonstrative modifier) and NPs with quantifiers (*all people* in the meaning *all these people*)
  - generic nouns co-refer with definite full NPs or pronouns, but not with other generics
- substitution: referring expression is replaced with another element
  Do you prefer the blue shirt or the red shirt?
  – I would like the red one.

Annotation Principles

UNIVERSITÄT DES SAARLANDES

- **Types of anaphora – pronouns and NPs:**
  - NPs: proper names (*Herr Almeida Freire*), nominal premodifiers, full NPs (with a definite article/demonstrative modifier) and NPs with quantifiers (*all people* in the meaning *all these people*)
  - generic nouns co-refer with definite full NPs or pronouns, but not with other generics
- substitution: referring expression is replaced with another element
  Do you prefer the blue shirt or the red shirt?
  – I would like the red one.
- ellipsis: referring expression is left out (implicit reference)
  ...if I take any one of these balls... and I count how many neighboring balls that there are around it, the answer's always twelve ____.

Annotation Principles

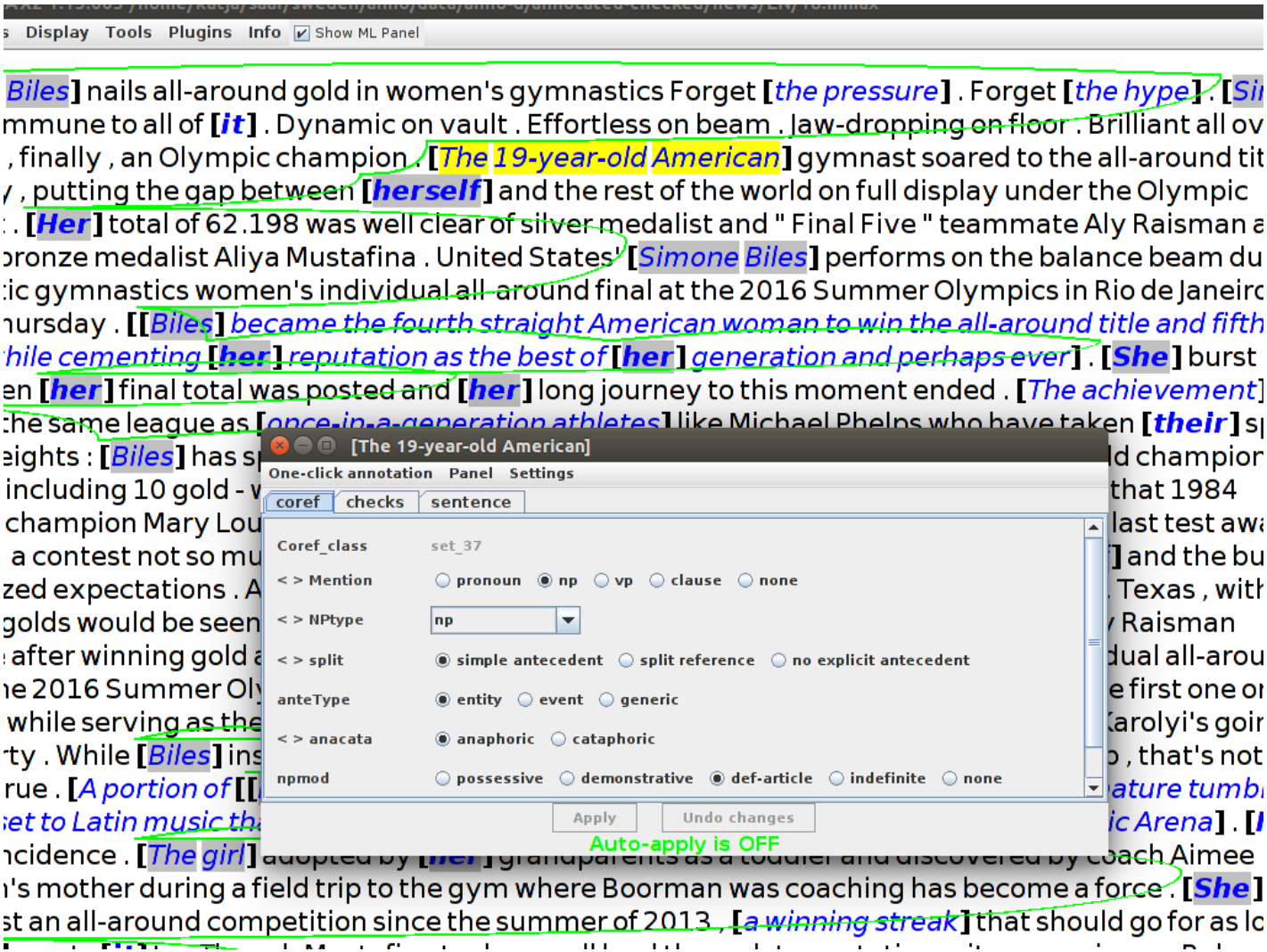
UNIVERSITÄT DES SAARLANDES

- **Types of anaphora – pronouns and NPs:**
  - NPs: proper names (*Herr Almeida Freire*), nominal premodifiers, full NPs (with a definite article/demonstrative modifier) and NPs with quantifiers (*all people* in the meaning *all these people*)
  - generic nouns co-refer with definite full NPs or pronouns, but not with other generics
- substitution: referring expression is replaced with another element
  Do you prefer the blue shirt or the red shirt?
  – I would like the red one.
- ellipsis: referring expression is left out (implicit reference)
  ...if I take any one of these balls... and I count how many neighboring balls that there are around it, the answer's always twelve ____.
- types of substitution/ellipsus: nominal, verbal and clausal

Annotation Process

UNIVERSITÄT DES SAARLANDES

manual with MMAX2 [Müller and Strube, 2006]



by highly experienced well-trained annotators with linguistic background

Data Selection

UNIVERSITÄT DES SAARLANDES

- existing resources: ParCor [Guillou et al., 2014]; DiscoMT2015 dataset [Hardmeier et al., 2015]
- exension:
  1. complete annotation of full coreference chains
  2. additional referring expressions to achieve full coreference chains

Data Selection

UNIVERSITÄT DES SAARLANDES

- existing resources: ParCor [Guillou et al., 2014]; DiscoMT2015 dataset [Hardmeier et al., 2015]
- exension:
  1. complete annotation of full coreference chains
  2. additional referring expressions to achieve full coreference chains
- additional data (to increase register/genre variety): WMT2017 data [Bojar et al., 2017].

Data Selection

UNIVERSITÄT DES SAARLANDES

- existing resources: ParCor [Guillou et al., 2014]; DiscoMT2015 dataset [Hardmeier et al., 2015]
- exension:
  1. complete annotation of full coreference chains
  2. additional referring expressions to achieve full coreference chains
- additional data (to increase register/genre variety): WMT2017 data [Bojar et al., 2017].

language	ParCor	DiscoMT	WMT news	TOTAL
English	31,971	39,764	10,644	82,379
German	30,305	37,452	10,593	78,350
TOTAL	62,276	77,216	21,237	**160,729**

Annotated Structures

UNIVERSITÄT DES SAARLANDES

	English	German	TOTAL
pronoun	4,650	4,269	8,919
np	2,485	2,611	5,096
vp	133	132	265
clause	335	312	647
total mentions	7,603	7,324	**14,927**

	English	German	TOTAL
number of chains	2,319	2,425	4,744
average chain length	2.94	2.81	2.87

Inter-annotator Agreement

UNIVERSITÄT DES SAARLANDES

- **evaluation**: 6,253 English and 5,975 German tokens
- **measure**: mention overlap
  and entity-based CEAF scores [Luo, 2005] b/n 2 annotations
- **scores** calculated with the CoNLL reference scorer
  implementation [Pradhan et al., 2014]

Inter-annotator Agreement

UNIVERSITÄT DES SAARLANDES

- **evaluation**: 6,253 English and 5,975 German tokens
- **measure**: mention overlap
  and entity-based CEAF scores [Luo, 2005] b/n 2 annotations
- **scores** calculated with the CoNLL reference scorer
  implementation [Pradhan et al., 2014]

	Precision	Recall	F-score
English			
mentions	89.20%	73.89%	80.71%
CEAF*e*	82.90%	67.13%	74.13%
German			
mentions	84.80%	69.76%	76.54%
CEAF*e*	72.53%	60.36%	65.88%

Summary and Outlook

parallel corpus with full **annotation of coreference**
$\Rightarrow$ valuable resource with a variety of uses:

Applications so far:

- discovery of inconsistencies presentation at ParCor-2018
- WMT pronoun translation evaluation [Guillou et al., 2018]

Future Applications

- help to study the **mechanisms involved in coreference translation** $\Rightarrow$ better understanding of the phenomenon
- a resource for **creating and evaluating coreference-aware MT** with no need in automatic coreference resolvers
- a **training and development** resource for **multilingual or monolingual coreference resolution** systems

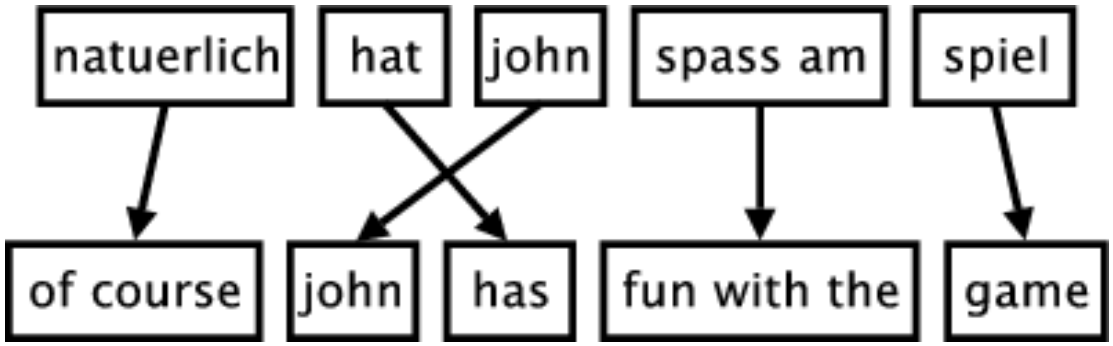
UNIVERSITÄT DES SAARLANDES

Example Study III: Alignment Discrepancies

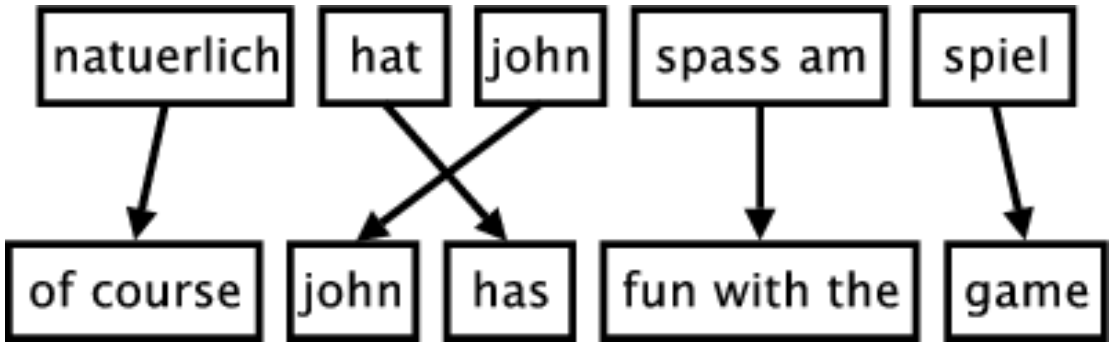
ongoing project

DiscoMT-Paper: [Lapshinova-Koltunski and Hardmeier, 2017]

Alignment Discrepancy: Definition



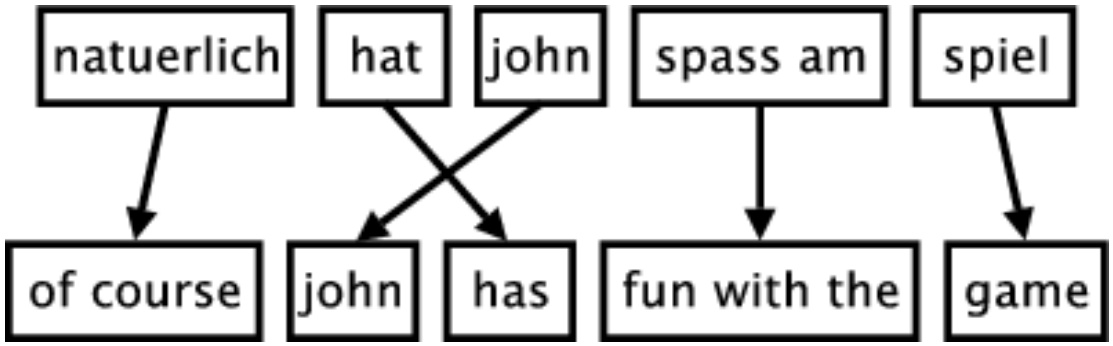
Alignment Discrepancy: Definition



Alignment Discrepancies

aligned sentence pairs with problems in word alignment

Alignment Discrepancy: Definition



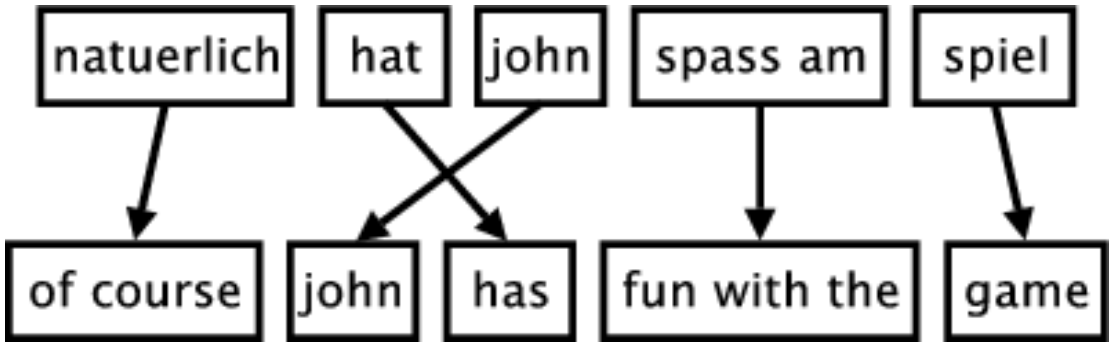
Alignment Discrepancies

aligned sentence pairs with problems in word alignment

Discourse-related

Educational researcher Benjamin Bloom, in 1984, posed what's called the 2 sigma problem, which he observed by studying three populations. 1984 veröffentlichte der Bildungsforscher Benjamin Bloom etwas, das '2-Sigma-Problem' heißt. Er beobachtete dies bei drei Populationen.

Alignment Discrepancy: Definition



Alignment Discrepancies

aligned sentence pairs with problems in word alignment

Discourse-related

Educational researcher Benjamin Bloom, in 1984, posed what's called the 2 sigma problem, which he observed by studying three populations. 1984 veröffentlichte der Bildungsforscher Benjamin Bloom etwas, das '2-Sigma-Problem' heißt. Er beobachtete dies bei drei Populationen.

Motivation: Discourse in (machine) translation

potential **elements of coreference chains** and contribute to the overall **coherence** and carry part of the **discourse information** in source / target languages

Discrepancy Source: Language Contrast

UNIVERSITÄT DES SAARLANDES

VS.

The demographic curves reveal that the welfare state can no longer be financed by the younger members of society. This does not mean that the country is descending into an unparalleled crisis ...

Die demographischen Kurven verraten, dass der Sozialstaat von den Jüngeren nicht mehr zu finanzieren ist. Damit versinkt das Land nicht in einer beinah unvergleichlichen Krise, wie manchmal behauptet wird.

Discrepancy Source: Translation Process

⇒

You want your employees to do what you ask them to do, and if they've done that, then they can do extra.

Sie erwarten von Ihren Angestellten, dass sie tun worum Sie sie gebeten haben, wenn sie die Aufgabe ausgeführt haben, können sie Zusätzliches tun.

Discrepancy Source: Alignment Error

UNIVERSITÄT DES SAARLANDES

Related Work: Coreference Projection

UNIVERSITÄT DES SAARLANDES

transformation pattern extraction:

- [Postolache et al., 2006]: patterns containing **heads** of the resulting referring expression **in the target** aligned with **heads** of **the source** referring expressions
- [Grishina and Stede, 2015]: apply a **direct projection algorithm** on parallel data to automatically produce coreference annotations for two target languages; describe a **number of problems**, when a referring expression is present in both source and target text but is not projected correctly

$\Rightarrow$ use manually annotated data

$\Rightarrow$ evidence is limited

$\Rightarrow$ no systematic description of discrepancies

Related Work: Contr./Translation Studies

UNIVERSITÄT DES SAARLANDES

- [Kunz and Steiner, 2012b, Kunz and Lapshinova-Koltunski, 2015]: **coreference relation** is shared across all languages, but the range of **referring expressions** is **different**
- [Meyer and Webber, 2013], [Becher, 2011b]: **explicitation** and **implicitation** – a source text does not contain linguistic markers that trigger some discourse relations, whereas its translation does, and vice verse some referring expressions can be more explicit than the others, **scale for the explicitness** of various referring expressions

We use **more data**
create a **new resource**
suggest **classification**
proceed **bottom-up**

Corpus Data

UNIVERSITÄT DES SAARLANDES

TED talks: Data from IWSLT 2015 MT Evaluation

Automatic annotations:

- token, lemma
- parts-of-speech and dependencies with Universal Dependency (UPOS und UD) treebank, [Nivre et al., 2015]
- sentence-/ word alignment (*mgiza*)

	sentences	tokens
English	214.889	3.940.079
German	227.649	3.678.503

194.370
parallel segments

Data Extraction

UNIVERSITÄT DES SAARLANDES

1. list of discourse-related structures (UPOS and gram.func from UD)

Data Extraction

UNIVERSITÄT DES SAARLANDES

1. list of discourse-related structures (UPOS and gram.func from UD)
2. personal pronouns und demonstratives (PRON and DET)

Data Extraction

UNIVERSITÄT DES SAARLANDES

1. list of discourse-related structures (UPOS and gram.func from UD)
2. personal pronouns und demonstratives (PRON and DET)
3. extraction of parallel segments using word alignment (1 : *N*)

Bsp `which DET-dobj` → `dies PRON-dobj`

Educational researcher Benjamin Bloom, in 1984, posed what's called the 2 sigma problem, which he observed by studying three populations.

1984 veröffentlichte der Bildungsforscher Benjamin Bloom etwas, das '2-Sigma-Problem' heißt. Er beobachtete dies bei drei Populationen.

Data Extraction

UNIVERSITÄT DES SAARLANDES

1. list of discourse-related structures (UPOS and gram.func from UD)
2. personal pronouns und demonstratives (PRON and DET)
3. extraction of parallel segments using word alignment (1 : *N*)

Bsp `which DET-dobj` → `dies PRON-dobj`

Educational researcher Benjamin Bloom, in 1984, posed what's called the 2 sigma problem, which he observed by studying three populations.

1984 veröffentlichte der Bildungsforscher Benjamin Bloom etwas, das '2-Sigma-Problem' heißt. Er beobachtete dies bei drei Populationen.

4. identification of non-aligned segments

General Observations

Overall ca. 11% alignment discrepancies

English sources (implicitation)

14 patterns (57,608 tokens)

German translations (explicitation)

26 patterns (69,851 tokens)

General Observations

Overall ca. 11% alignment discrepancies

English sources (implicitation)

14 patterns (57,608 tokens)

German translations (explicitation)

26 patterns (69,851 tokens)

Quantitative und qualitative analysis:

1. Which are the most frequent structures in the English sources and the German translations?
   most frequent: DET-det, PRON-subj und PRON-nmod oder PRON-dobj; others vary in English und German
2. Are there language-specific patterns?
   occur in one list only: 6 out of 14 English, 25 out of 26 German

Observations: DET-det

UNIVERSITÄT DES SAARLANDES

Generics

You know, it's just like the hail goes out and people are ready to help.
Es ist einfach so, jemand ruft um Hilfe, und die Leute stehen zur Hilfe bereit.

Observations: DET-det

UNIVERSITÄT DES SAARLANDES

Generics

You know, it's just like the hail goes out and people are ready to help.
Es ist einfach so, jemand ruft um Hilfe, und die Leute stehen zur Hilfe bereit.

Explicitation

You would expect it to be cheesy, but it 's not.
Man könnte annehmen, dass so etwas kitschig ist, aber das ist nicht der Fall.

Observations: DET-det

UNIVERSITÄT DES SAARLANDES

Generics

You know, it's just like the hail goes out and people are ready to help.
Es ist einfach so, jemand ruft um Hilfe, und die Leute stehen zur Hilfe bereit.

Explicitation

You would expect it to be cheesy, but it 's not.
Man könnte annehmen, dass so etwas kitschig ist, aber das ist nicht der Fall.

Implicitation

Secondly, there had to be an acceptance that we were not going to be able to use all of this vacant land in the way that we had before and maybe for some time to come.
Zweitens musste es eine Übereinkunft geben, dass wir das gesamte brachliegende Land nicht wie vorher nutzen können würden...

Observations: PRON-nsub

UNIVERSITÄT DES SAARLANDES

Finite vs. non-finite constructions (*ing*-forms vs. Partizip I)

A polar bear swimming in the Arctic, by Paul Nicklen.
Ein Eisbär, der in der Arktis schwimmt, aufgenommen von Paul Nicklen.

Observations: PRON-nsub

UNIVERSITÄT DES SAARLANDES

Finite vs. non-finite constructions (*ing*-forms vs. Partizip I)

A polar bear swimming in the Arctic, by Paul Nicklen.
Ein Eisbär, der in der Arktis schwimmt, aufgenommen von Paul Nicklen.

Impersonalisation

If you have fluid with no wall to surround it and keep pressure up, you have a puddle.
Eine Flüssigkeit ohne eine Wand, die sie umgibt und den Druck aufrechterhält, ist eine Pfütze.

Observations: PRON-nsub

UNIVERSITÄT DES SAARLANDES

Finite vs. non-finite constructions (*ing*-forms vs. Partizip I)

A polar bear swimming in the Arctic, by Paul Nicklen.
Ein Eisbär, der in der Arktis schwimmt, aufgenommen von Paul Nicklen.

Impersonalisation

If you have fluid with no wall to surround it and keep pressure up, you have a puddle.
Eine Flüssigkeit ohne eine Wand, die sie umgibt und den Druck aufrechterhält, ist eine Pfütze.

Syntax / Cohesion

Some days it goes up and some days it doesn't go up.
An manchen Tagen geht er hoch und an manchen Tagen nicht.

Back to Theories

UNIVERSITÄT DES SAARLANDES

explanation attempt: explicitation

1. [Blum-Kulka, 1986]: “explicitation hypothesis”
2. [Klaudy, 2008]: Classification:
   (1) obligatory; (2) optional; (3) pragmatic;
   (4) translation-specific
3. [Zufferey and Cartoni, 2014]: Explicitation scale:
   no $\rightarrow$ light $\rightarrow$ strong

Back to Theories

UNIVERSITÄT DES SAARLANDES

explanation attempt: explicitation

1. [Blum-Kulka, 1986]: "explicitation hypothesis"
2. [Klaudy, 2008]: Classification:
   (1) obligatory; (2) optional; (3) pragmatic;
   (4) translation-specific
3. [Zufferey and Cartoni, 2014]: Explicitation scale:
   no $\rightarrow$ light $\rightarrow$ strong

(1) + (2) useful for our observations?

- **obligatory** (systemic differences):
  WHEN no explicitation **THEN** non-grammatical constructions
- **optional** (stylistic and genre-/register-specific differences):
  WHEN no explicitation **THEN** unnatural constructions

Back to Theories

explanation attempt: constrative pragmatics/grammar

UNIVERSITÄT DES SAARLANDES

corpus data: addressee reference ⇒ impersonalisation

- **finite** → **non-finite** form, *you* → *man*:
  It means that, if you have T.B., you have more chance that you won't be detected...
  Das heißt, wenn man T.B. hat, ist die Wahrscheinlichkeit, nicht diagnostiziert zu werden...
- **active** → **passive**:
  And this starts conversation because the visual is created in such a way where you can defend two answers.
  Und das startet eine Konversation denn das Bild ist so konstruiert, dass zwei Antworten verteidigt werden können.

Back to Theories

explanation attempt: constrative pragmatics/grammar

UNIVERSITÄT DES SAARLANDES

corpus data: addressee reference ⇒ impersonalisation

- **finite** → **non-finite** form, *you* → *man*:
  It means that, if you have T.B., you have more chance that you won't be detected...
  Das heißt, wenn man T.B. hat, ist die Wahrscheinlichkeit, nicht diagnostiziert zu werden...
- **active** → **passive**:
  And this starts conversation because the visual is created in such a way where you can defend two answers.
  Und das startet eine Konversation denn das Bild ist so konstruiert, dass zwei Antworten verteidigt werden können.

1. contrastive pragmatics
   [House, 2014, Becher, 2011a, Kranich, 2016]

Back to Theories

explanation attempt: constrative pragmatics/grammar

UNIVERSITÄT DES SAARLANDES

corpus data: addressee reference ⇒ impersonalisation

- **finite** → **non-finite** form, *you* → *man*:
  It means that, if you have T.B., you have more chance that you won't be detected...
  Das heißt, wenn man T.B. hat, ist die Wahrscheinlichkeit, nicht diagnostiziert zu werden...
- **active** → **passive**:
  And this starts conversation because the visual is created in such a way where you can defend two answers.
  Und das startet eine Konversation denn das Bild ist so konstruiert, dass zwei Antworten verteidigt werden können.

1. contrastive pragmatics
   [House, 2014, Becher, 2011a, Kranich, 2016]
2. contrastive grammar
   [Hawkins, 1986, Doherty, 2003, Steiner, 2012, Fischer, 2013]

Summary and Outlook

UNIVERSITÄT DES SAARLANDES

used parallel corpus for:

Applications so far:

- **quantitatively** describe **alignment discrepancies** between English-German discourse-related phenomena from a language contrastive perspective
- interesting observations **different from traditional grammar**, e.g. those on the article use with generics or non-finite constructions

Future Work

- more cases in the future
- information on **systematic discrepancies** for alignment improvement
- **explicitation types** more (systematic) description here
- **applications**, e.g. improve models to avoid the overgeneration

UNIVERSITÄT DES SAARLANDES

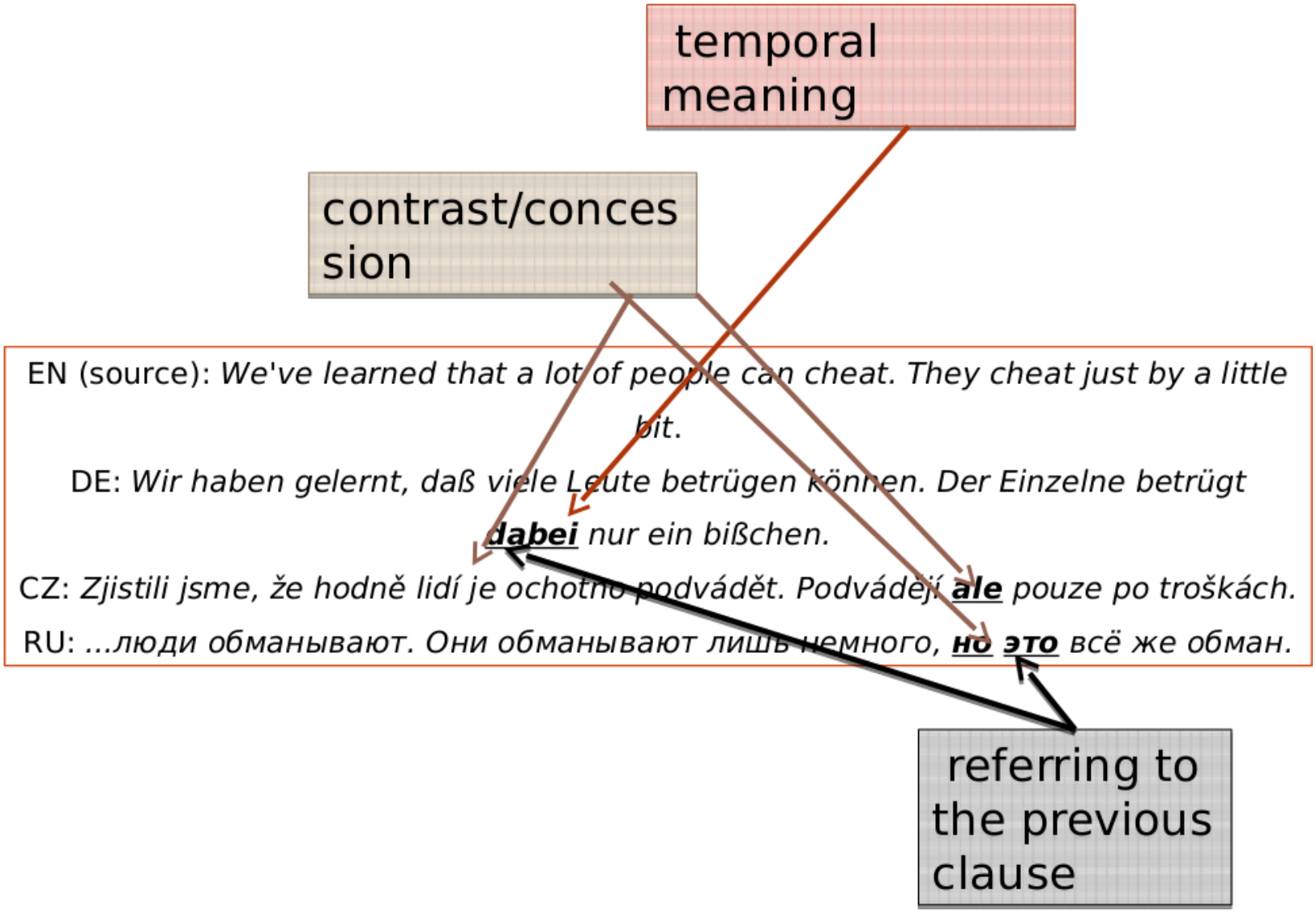
Example Study IV: Pronominal adverbs in DE and their equivalents in EN, CZ, RU

ongoing project (follow-up Textlink EU Cost)

Dialogue-2018 paper:

[Nedoluzhko and Lapshinova-Koltunski, 2018]

Pronominal Adverbs in German



Pronominal Adverbs in German

UNIVERSITÄT DES SAARLANDES

Observations

- Pronominal adverbs are frequent in DE and rarely occur in the other languages under analysis
- DE pronominal adverbs are ambiguous
- There is a great variation in their equivalents in the corresponding languages
- German seems to be more explicit (if pronominal adverbs are concerned)

Aims and Goals

UNIVERSITÄT DES SAARLANDES

- analyse multilingual parallel data (DE, EN, CZ, RU) – TED/WMT17
- discover and analyse various means that correspond to DE pronominal adverbs in EN, CZ and RU
- describe their functions and usage preferences trying to find systematicity in their usage
- find out what impact may translation process have on the choice in languages?

Results

UNIVERSITÄT DES SAARLANDES

Functions of DE pronominal adverbs are genre-specific

- **anaphoric function** of pronominal adverbs more frequent in TED (popsci talks):

  DE: *Ein brilliantes spätes Tackling von Marcus Watson... sicherte den Gewinn – und damit die Silbermedaille*. (in this way)
  EN: *A brilliant late tackle from Marcus Watson... secured the win – and ultimately the silver medal.*

- **correlative function** is predominating in WMT17 (news):

  DE: *Die Bedeutung des Wortes "Patient" habe nichts damit zu tun, Ratschläge zu geben.* (with this)
  EN: *The word "patient" doesn't mean to make suggestions*.

Results

UNIVERSITÄT DES SAARLANDES

Transformation patterns reflect language contrasts

- verbal valency: *hören Sie damit auf* – preposition *mit* is obligatory!
- syntactic constructions:

  EN gerund vs. DE, CZ and RU construction with a correlative:
  EN: *I dream of bringing*
  DE: *Ich träume davon ... zu bringen*

Translation process: explicitation/implicitation

- zero $\Leftrightarrow$ discourse element
- less explicit $\Leftrightarrow$ more explicit

  EN: *I didn't know if I could do that and preserve my art.*
  DE: *Ich wußte nicht, ob ich das bewerkstelligen, und dabei meine Kunst bewahren konnte.*
  CZ: *Nevěděla jsem, jestli to dokážu udělat a zachovat přitom své umění.*

Summary and Outlook

used parallel corpus with no **annotation of coreference**
$\Rightarrow$ analyse coreference-related phenomena

Applications so far:

- discovery of interesting language contrasts in use
- phenomena of translations process (contrast-induced)

Future Work

- more cases (200 manually analysed so far)
- automatisation of categories?
- a resource with a variety of uses

Two Approaches

UNIVERSITÄT DES SAARLANDES

UNIVERSITÄT DES SAARLANDES

Thank you very much for your attention!

Dziękuję ... bardzo za ... uwagę!
Dziękuję Państwu bardzo za ... uwagę!
Dziękuję ... bardzo za Pańską uwagę!

e.lapshinova@mx.uni-saarland.de

Becher, V. (2011a).

Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts.
PhD thesis, Universität Hamburg.

Becher, V. (2011b).

When and why do translators add connectives? a corpus-based study.
Target, 23.

Blum-Kulka, S. (1986).

Shifts of cohesion and coherence in translation.
In House, J. and Blum-Kulka, S., editors, *Interlingual and intercultural communication*, pages 17–35. Gunter Narr, Tübingen.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017).
Findings of the 2017 conference on machine translation (wmt17).
In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Doherty, M. (2003).

Language Processing in Discourse: A Key to Felicitous Translation.
Routledge Studies in Germanic Linguistics. Taylor & Francis.

Fischer, K. (2013).

Satzstrukturen im Deutschen und Englischen. Typologie und Textrealisierung.
Akademie Verlag, Berlin.

Grishina, Y. and Stede, M. (2015).

Knowledge-lean projection of coreference chains across languages.
In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora, Beijing, China*, page 14.

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (2018).

A pronoun test suite evaluation of the english–german mt systems at wmt 2018.
In *Proceedings of the Third Conference on Machine Translation at EMNLP-2018*. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014).

ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT.
In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland.

Hardmeier, C., P., Nakov, S., Stymne, J., Tiedemann, Y., Versley, and Cettolo, M. (2015).

Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation.
In *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), EMNLP 2015*, pages 1–16, Lisbon, Portugal.

Hawkins, J. A. (1986).

A Comparative Typology of English and German.
Croom Helm, London and Sydney.

House, J. (2014).

Translation Quality Assessment. Past and Present.
Routledge.

Klaudy, K. (2008).

Explicitation.
In Baker, M. and Saldanha, G., editors, *Routledge Encyclopedia of Translation Studies*, pages 104–108. Routledge, London & New York, 2 edition.

Kranich, S. (2016).

Contrastive Pragmatics and Translation. Evaluation, Epistemic Modality and Communicative Style in English and German, volume 261 of *Pragmatics and Beyond, New Series*.
John Benjamins, Amsterdam.

Kunz, K. and Lapshinova-Koltunski, E. (2015).

Cross-linguistic analysis of discourse variation across registers.
Special Issue of Nordic Journal of English Studies, 14(1):258–288.

Kunz, K. and Steiner, E. (2012a).

Towards a comparison of cohesive reference in English and German: System and text.
In Taboada, M., Suárez, S. D., and Álvarez, E. G., editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*. Equinox, London.

Kunz, K. and Steiner, E. (2012b).

Towards a comparison of cohesive reference in english and german: System and text.
In Taboada, M., Suárez, S. D., and Álvarez, E. G., editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*. Equinox, London.

Lapshinova, E., Herdmeier, C., and Krielke, P. (2018).

ParCorFull: a Parallel Corpus Annotated with Full Coreference.
In *Proceedings of LREC-2018*, Miyazaki, Japan. ELDA.

Lapshinova-Koltunski, E. and Hardmeier, C. (2017).

Discovery of discourse-related language contrasts through
alignment discrepancies in english-german translation.
In *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT 2017) at EMNLP-2017*, Copenhagen, Denmark.

Lapshinova-Koltunski, E. and Kunz, K. (2014).

Annotating cohesion for multillingual analysis.
In *Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 57–64, Reykjavik, Iceland.

Luo, X. (2005).

On coreference resolution performance metrics.
In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada.

Meyer, T. and Webber, B. (2013).

Implicitation of discourse connectives in (machine) translation.
In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Müller, C. and Strube, M. (2006).

Multi-level annotation of linguistic data with MMAX2.
In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Nedoluzhko, A. and Lapshinova-Koltunski, E. (2018).

Pronominal adverbs in german and their equivalents in english, czech and russian: Evidence from the parallel corpus.
In *Computational Linguistics and Intellectual Technologies: Proceedings of the 24th International Conference "Dialogue-21"*, Moscow. RSUH.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015).
Universal dependencies 1.0.
LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Postolache, O., Cristea, D., and Orasan, C. (2006).

Tranferring coreference chains through word alignment.
In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014).

Scoring coreference partitions of predicted mentions: A reference implementation.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland.

Steiner, E. (2012).

Introduction.
In *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*, pages 1–17. de Gruyter, Berlin, New York.

Zufferey, S. and Cartoni, B. (2014).
A multifactorial analysis of explicitation in translation.
Target, 26(3):361–384.