Statistical Insights into Cohesion

Contrasting English and German across Modes

Ekaterina Lapshinova-Koltunski and José Martínez-Martínez

FR4.6, UdS, Saarbrücken

May 25, 2016

Background

Research Project

GECCo: German-English Contrasts in Cohesion

supported by the DFG

Project Team:

- Kerstin Kunz
- Ekaterina Lapshinova-Koltunski
- Erich Steiner
- Jose Manuel Martinez
- Katrin Menzel

Acknowledgement: Stefania Degaetano-Ortlieb, Marilisa Amoia

FR 4.6 Applied Linguistics, Interpreting and Translation Studies

www.gecco.uni-saarland.de

Overview

1. Defining Concept
2. Research Agenda and Methodology
3. Analyses
   - Dis/similarities between variables and features
   - Distinctive Features
4. Conclusions

Defining Concept

Cohesive Phenomena

Cohesion is an important component of effectively organised and meaningful **discourse**, as the **message** being communicated in discourse is not just a set of clauses, but forms a **unified, coherent whole**

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, …)

Several studies have shown that two of the factors affecting regret are how much one feels personal responsibility for the result and how easy it is to imagine a better alternative. The availability of choice obviously exacerbates both these factors . When you have no options, what can you do? You will feel disappointment, maybe; regret, no. With no options, you just do the best you can. But with many options, the chances increase that a really good one is out there, and you may well feel that you ought to have been able to find it.

Mehreren Studien zufolge wird das Gefühl der Reue zum einen stärker, je mehr man sich für das Resultat persönlich verantwortlich fühlt, und zum anderen, je leichter man sich eine bessere Alternative vorstellen kann. Ein Auswahlangebot verschlimmert offensichtlich beide Faktoren. Was kann man schon groß anstellen, wenn man keine Wahl hat? Vielleicht ist man enttäuscht, aber Reue empfindet man nicht. Wenn es hingegen viele Optionen gibt, wächst das Risiko, dass man meint, eine besonders gute übersehen zu haben, und dies nun bereut.

Research Agenda and Methodology

Research Questions

1. How cohesive are the texts in English and German / in spoken and written texts?
2. How strong are cohesive relations?
3. Which semantic relations are generally expressed and which cohesive devices are preferred over others?
4. How much cohesive variation is there in one language as compared to the other?

Methodology

- compare EO vs. GO
  (Hawkins, 1986; König&Gast, 2012; Königs, 2011, etc.)
- compare spoken vs. written
  (Mair, 2006; Leech et al., 2009)
- compare registers
  (Hansen-Schirra et al., 2012; Neumann, 2013)
  - in terms of number of cohesive devices
  - in terms of number of chains, length of chains

→ **corpus-based analysis**:

- ▶ define operationalisations
- ▶ extract instances/frequencies from corpus
- ▶ evaluate frequencies statistically

GECCCOH

subcorpora	registers
written	imported from CroCo*
EO 🇬🇧 GO 🇩🇪	FICTION, ESSAY, INSTR, POPSCI, SHARE, SPEECH, TOU, WEB
spoken	collected at FR4.6, UdS**
EO-SPOKEN 🇬🇧 GO-SPOKEN 🇩🇪	INTERVIEW, ACADEMIC, FORUM, TALKSHOW, MEDCONSULT, SERMON

GECCo annotation levels

1) word: ⇒ *word, lemma, pos*

2) chunk:⇒ *sentences, syntactic chunks, clauses, cohesion*

3) text: ⇒ *registers*

4) extralinguistic: ⇒ *register analysis, speaker information*

* cf. (Hansen-Schirra et al., 2012)

** cf. (Lapshinova et al., 2012)

GECCCOH

CQP= Corpus Query Processor, cf. (Evert 2005)

Positional Attributes:	word pos lemma
Structural Attributes:	NP_gf VP_gf PP_gf sentence reference_type reference_function conjunction_type conjunction_function text text_register

Annotation of Cohesion

(Lapshinova & 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>
```

ANALYSES

Types of Analyses

1. Dis/similarities between variables and features:
   Correspondence Analysis (CA),
   cf. (Baayen, 2008) & (Greenacre, 2010)

2. Features, distinctive for each variable:
   Classification with Support Vector Machines (SVM),
   cf. (Vapnik & Chervonenkis, 1974; Joachims, 1998)

cf. (Kunz et al. forthcoming)

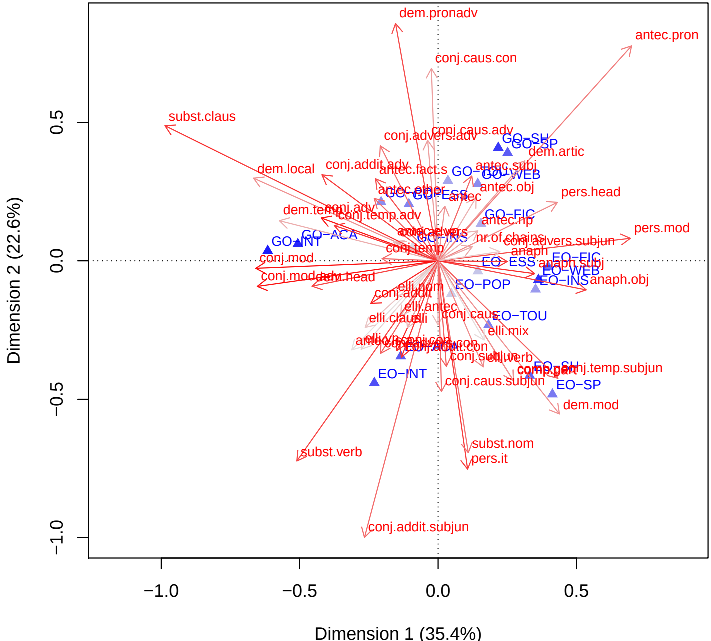
Features and their Combinations

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			

Correspondence Analysis

- **Input:** frequencies of cohesive devices across registers and languages
- **Output:** a multi-dimensional plot, in which the co-related variables are scattered
  - **arrows** for the observed feature frequencies
  - **points** for registers across languages
- **Interpretation:**
  - the larger the differences between subcorpora, the further apart **they** are on the map → dissimilar categories of **coh.dev.** are further apart
  - the position of the **points** in relation to the **arrows** indicates the relative importance of a feature for a register.
  - the length of the **arrows** indicates how pronounced a particular feature is
  - the **arrows** pointing in the direction of an axis indicate a high contribution to the respective dimension

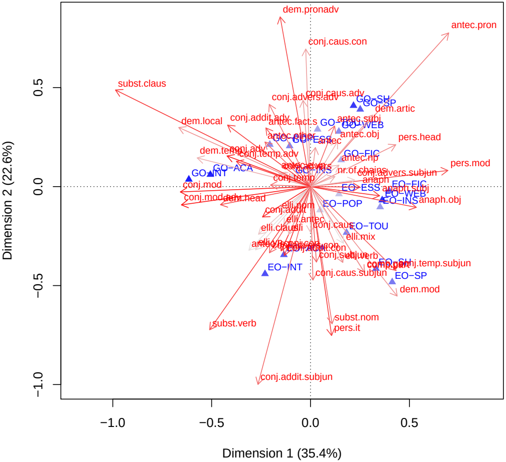
CA: all features



CA: all features

→ groups of subcorpora:

- **x-axis**: clear differences between registers
  - **features**: conj. relations and coreference

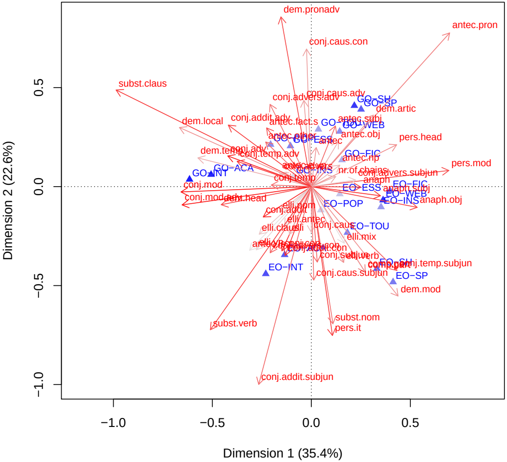


CA: all features

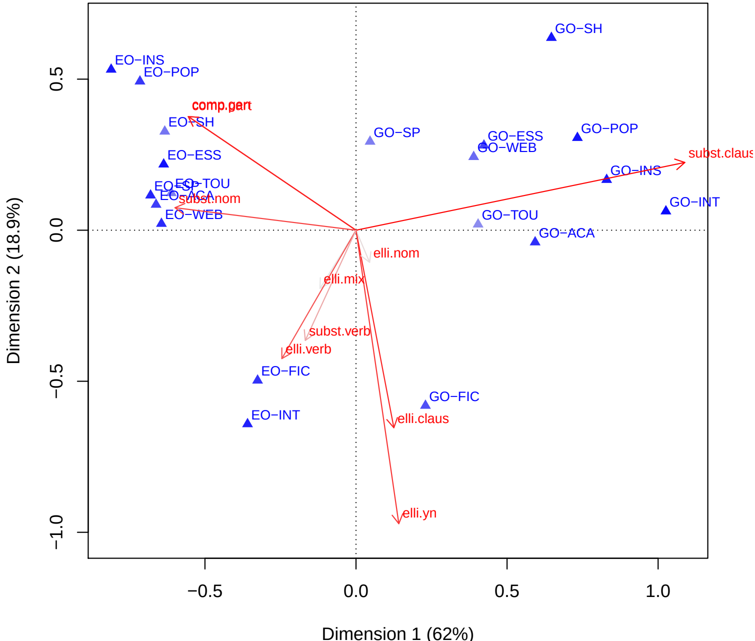
→ groups of subcorpora:

- **y-axis**: differences languages

- **features**: dem.pronadv vs. pers.it conj: causal vs. addit. and con. vs. subj)



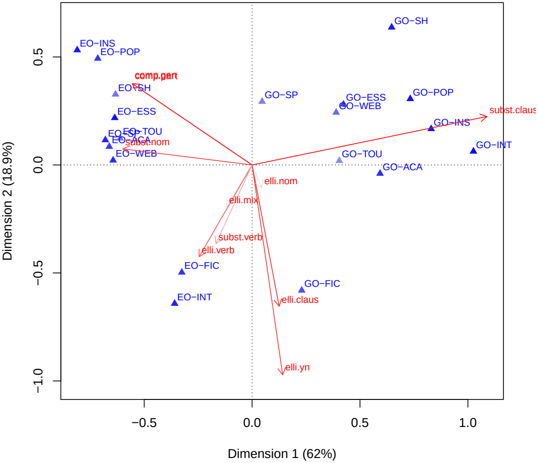
CA: similarity



CA: similarity

→ groups of subcorpora:

- **x-axis**: clear differences between languages
  - **features**: claus. and nom. substitution
- **y-axis**: registers
  - **features**: ellipses



Text Classification Technique

⇒ identify distinctive features

- individual texts are classified into classes
- classes are intrinsically defined
- pairwise classification: a set of one-versus-one classifier is built due to multiple classes
- Support Vector Machines* with 10-folds cross-validation
- SMO (Sequential minimal optimization) SVM with linear kernal

*(Vapnik & Chervonenkis, 1974; Joachims, 1998)

Classification: Language

	Precision	Recall	F-Measure
GO	99.3	98.6	99.0
EO	98.5	99.2	98.9
Weight.Av.	98.9	98.9	98.9

GO		EO	
4.1984	coref:dem-pronadv	1.0173	coref:comp-general
2.0487	conj:adversat-adverb	1.9675	coref:dem-mod
1.4926	conj:causal-adverb	1.6618	subst:nom
1.4716	subst:claus	1.6057	coref:pers-it
1.1850	coref:dem-local	1.4476	conj:causal-subjun
1.1568	coref:dem-artic	1.3708	conj:temp-subjun
1.0611	conj:addit-adverb	1.1606	subst-verb
1.0585	conj:temp-adverb	1.0173	coref:comp-particular
0.9209	conj:modal-adverb	0.9146	conj:adversat-conn
0.9135	conj:adversat-subjun	0.8751	coref:pers-mod

Classification: Mode

	Precision	Recall	F-Measure
spoken	97.7	91.3	94.4
written	98.3	99.6	98.9
Weight.Av.	98.2	98.2	98.2

written		**spoken**	
0.8866	conj:temp-subjun	1.3347	conj:modal-adverb
0.8543	coref:pers-mod	1.1550	coref:pers-it
0.7204	coref:obj	1.1275	subst:verb
0.7177	elli:verb	1.0998	coref:dem-head
0.6995	conj:causal-adverb	0.9904	conj:adversat-conn
0.5736	conj:causal-conn	0.8789	conj:addit-conn
0.5107	antec:subj	0.6856	conj:addit-subjun
0.4760	conj:adversat-adverb	0.6239	subst:nom
0.4485	conj:adversat-subjun	0.4989	antec:other
0.4120	antec:obj	0.4709	antec:event-vp

CONCLUSIONS

Research Questions

1. How cohesive are the texts in English and German / in spoken and written dimensions?
   - German = English, spoken > written
2. How strong are cohesive relations?
   - German: wider scope, stronger specification, more focused vs. English
   - spoken: wider scope, weaker specification, more focused vs. written
3. Which semantic relations are generally expressed and which cohesive devices are preferred over others?
   - German: logico-sem. (contrast and manner), identity
   - English: identity, similarity
   - spoken: similarity, logico-sem. (explanation)
   - written: identity, contrast and manner
4. How much cohesive variation is there in one language as compared to the other?
   - German > English

Thank you!

Questions?

Contact:
e.lapshinova@mx.uni-saarland.de
j.martinez@mx.uni-saarland.de
Information:
www.gecco.uni-saarland.de

References

- Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: CUP.
- S. Evert, 2005. The CQP Query Language Tutorial. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, April. CWB version 2.2.b90.
- Fabricius-Hansen, C. (1996). Informational density: A problem for translation theory. Linguistics 34: 521-565 (special issue).
- Greenacre, M. (2010). Correspondence Analysis in Practice, Second Edition. CRC Press.
  Halliday, M.A.K. (1989). Spoken and Written Language. Oxford: Oxford Universi- ty Press
- Halliday, M.A.K. and R. Hasan. (1976). Cohesion in English. Longman, London.
- Hansen-Schirra, Silvia, Neumann, Stella and Steiner, Erich (2012). Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English - German. Series Text, Translation, Computational Processing. Berlin, New York: Mouton de Gruyter.
- Hansen-Schirra, S., S. Neumann, and E. Steiner (2007). Cohesion and Explicitation in an English-German Translation Corpus. In: Languages in Con- trast 7(2): 241-265.
- Hawkins. John A. 1986. A comparative typology of English and German. Unifying the contrasts. London etc. Croom Helm.
- House, J. (1997). Translation Quality Assessment. Tübingen: Narr.
- H. Kermes and S. Evert (2002). YAC – A Recursive Chunker for Unrestricted German Text. In Manuel Gonzalez Rodriguez and Carmen PazSuarez Araujo, (eds). In Proceedings of the Third International Conference on Language Resources and Evaluation, pp 1805-1812.
- H. Kermes (2003). Off-line (and On-line) Text Analysis for Computational Lexicography. Ph.D. thesis, Universität Stuttgart.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine Learning: ECML98, 137-142.
- König, E. & V. Gast (2012). Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik. Berlin: Erich Schmidt Verlag. [3rd, extended edition].
- Königs, K. (2011). Übersetzen Englisch-Deutsch, Lernen mit System. Oldenbourg Verlag.
- Kunz, K. (2010). Variation in English and German Coreference. Frankfurt/ Main: Peter Lang.

References

- Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K. & Steiner, E. (submitted). GECCo - an empirically-based comparison of English-German cohesion. In De Sutter, G. and Delaere, I. and Lefer, M.-A. (eds.). New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies. TILSM series. Mouton de Gruyter.
- Lapshinova-Koltunski, E. & K. Kunz (2014). Detecting Cohesion: semi-automatic annotation procedures. In: Proceedings of Corpus Linguistics. Lancaster, UK, July.
- Lapshinova-Koltunski, E., K. Kunz and M. Amoia (2012). Compiling a Multilingual Corpus. In Mello, H., and M. Pettorino (eds). Proceedings of the VIIthGSCP-2012. Speech and Corpora. Firenze: Firenze University Press.
- Levy, R. and T. F. Jaeger (2007). Speakers Optimize Information Density Through Syntactic Reduction. In: Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS).
- M.M. Louwerse & A.C. Graesser (2005). Coherence in discourse. In P Strazny (ed.), Encyclopedia of linguistics, pp 216-218. Fitzroy Dearborn, Chicago.
- Leech, G., M. Hundt, C. Mair, & N. Smith (2009). Change in Contemporary English. A
- Mair, C. (2006). Twentieth-Century English. History, variation and standardization.
- Müller, C. and M. Strube (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee (eds)., Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, Peter Lang, Frankfurt a.M., Germany.
- Neumann, Stella (2013). Cross-linguistic Register variation. Mouton De Gruyter.
- Steiner, E. (2005). Some properties of texts in terms of ’information distribution across languages’. Languages in Contrast 5(1): 49-72.
- Venables, W.N. and D.M. Smith (2010). An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Electronic edition.
- Vapnik, V.N. and A.J. Chervonenkis (1974). Theory of Pattern Recognition. Nauka, Moscow.