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
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)

<table>
<thead>
<tr>
<th>Cohesive conjunction</th>
<th>Meaning relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>X. <strong>But</strong>/ <strong>And</strong>/ <strong>However</strong> Y</td>
<td>Logico-semantic relations (addition, contrast, cause, ...)</td>
</tr>
<tr>
<td>Coreference</td>
<td>identity</td>
</tr>
<tr>
<td><em>An option ... it/ this option</em></td>
<td></td>
</tr>
<tr>
<td>Substitution</td>
<td>Type reference/ comparison</td>
</tr>
<tr>
<td><em>Many options ... a good one</em></td>
<td></td>
</tr>
<tr>
<td>Ellipsis</td>
<td></td>
</tr>
<tr>
<td><em>You will feel disappointment. [ ]</em></td>
<td></td>
</tr>
<tr>
<td><em>Maybe.</em></td>
<td></td>
</tr>
<tr>
<td><em>Many options ... a good [ ].</em></td>
<td></td>
</tr>
<tr>
<td>Comparative Reference</td>
<td></td>
</tr>
<tr>
<td><em>One option ... another/ better option</em></td>
<td></td>
</tr>
</tbody>
</table>

### Meaning relations

- **Identity**
- **Type reference/comparison**
- **Logico-semantic relations** (addition, contrast, cause, ...)

---

**References:**
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.
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

<table>
<thead>
<tr>
<th>subcorpora</th>
<th>registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>written</td>
<td>imported from CroCo*</td>
</tr>
<tr>
<td>EO</td>
<td>FICTION, ESSAY, INSTR,</td>
</tr>
<tr>
<td>GO</td>
<td>POPSCI, SHARE, SPEECH, TOU, WEB</td>
</tr>
<tr>
<td>spoken</td>
<td>collected at FR4.6, UdS**</td>
</tr>
<tr>
<td>EO-SPOKEN</td>
<td>INTERVIEW, ACADEMIC,</td>
</tr>
<tr>
<td>GO-SPOKEN</td>
<td>FORUM, TALKSHOW, MEDCONSULT, SERMON</td>
</tr>
</tbody>
</table>

GECCCo 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)

<table>
<thead>
<tr>
<th>Positional Attributes:</th>
<th>word</th>
<th>pos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural Attributes:</td>
<td>NP_gf</td>
<td>VP_gf</td>
<td>PP_gf</td>
</tr>
<tr>
<td></td>
<td>sentence</td>
<td>reference_type</td>
<td>reference_function</td>
</tr>
<tr>
<td></td>
<td>conjunction_type</td>
<td>conjunction_function</td>
<td></td>
</tr>
<tr>
<td></td>
<td>text</td>
<td>text_register</td>
<td></td>
</tr>
</tbody>
</table>
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

<daraus>
<hier>
<grössere>
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

<table>
<thead>
<tr>
<th>COREFERENCE</th>
<th>SUBSTITUION</th>
<th>CONJUNCTION</th>
<th>ELLIPSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>antecedent-np, antecedent-pronominal, antecedent-fact-s, antecedent-event-vp, antecedent-is-a, antecedent-other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all anaphors</td>
<td></td>
<td>conj-addit, conj-adversat, conj-causal, conj-temp, conj-modal</td>
<td>all elli</td>
</tr>
<tr>
<td>antecedent-subj, antecedent-obj, anaphors-subj, anaphors-obj</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
Analyses Dis/similarities between variables and features

CA: similarity

→ groups of subcorpora:
  • **x-axis**: clear differences between languages
  - **features**: claus. and nom. substitution
  • **y-axis**: registers
  - **features**: ellipses
Analyses Distinctive Features

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 kernel

*(Vapnik & Chervonenkis, 1974; Joachims, 1998)
## Classification: Language

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>99.3</td>
<td>98.6</td>
<td>99.0</td>
</tr>
<tr>
<td>EO</td>
<td>98.5</td>
<td>99.2</td>
<td>98.9</td>
</tr>
<tr>
<td>Weight.Av.</td>
<td>98.9</td>
<td>98.9</td>
<td>98.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GO</th>
<th>EO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Feature</td>
<td>Score</td>
</tr>
<tr>
<td>4.1984</td>
<td>coref:dem-pronadv</td>
<td>1.0173</td>
</tr>
<tr>
<td>2.0487</td>
<td>conj:adversat-adverb</td>
<td>1.9675</td>
</tr>
<tr>
<td>1.4926</td>
<td>conj:causal-adverb</td>
<td>1.6618</td>
</tr>
<tr>
<td>1.4716</td>
<td>subst:claus</td>
<td>1.6057</td>
</tr>
<tr>
<td>1.1850</td>
<td>coref:dem-local</td>
<td>1.4476</td>
</tr>
<tr>
<td>1.1568</td>
<td>coref:dem-artic</td>
<td>1.3708</td>
</tr>
<tr>
<td>1.0611</td>
<td>conj:addit-adverb</td>
<td>1.1606</td>
</tr>
<tr>
<td>1.0585</td>
<td>conj:temp-adverb</td>
<td>1.0173</td>
</tr>
<tr>
<td>0.9209</td>
<td>conj:modal-adverb</td>
<td>0.9146</td>
</tr>
<tr>
<td>0.9135</td>
<td>conj:adversat-subjunct</td>
<td>0.8751</td>
</tr>
</tbody>
</table>

May 25, 2016
### Classification: Mode

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoken</td>
<td>97.7</td>
<td>91.3</td>
<td>94.4</td>
</tr>
<tr>
<td>written</td>
<td>98.3</td>
<td>99.6</td>
<td>98.9</td>
</tr>
<tr>
<td>Weight.Av.</td>
<td>98.2</td>
<td>98.2</td>
<td>98.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>written</th>
<th>spoken</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8866</td>
<td>conj:temp-subjun</td>
<td>1.3347</td>
</tr>
<tr>
<td>0.8543</td>
<td>coref:pers-mod</td>
<td>1.1550</td>
</tr>
<tr>
<td>0.7204</td>
<td>coref:obj</td>
<td>1.1275</td>
</tr>
<tr>
<td>0.7177</td>
<td>elli:verb</td>
<td>1.0998</td>
</tr>
<tr>
<td>0.6995</td>
<td>conj:causal-adverb</td>
<td>0.9904</td>
</tr>
<tr>
<td>0.5736</td>
<td>conj:causal-conn</td>
<td>0.8789</td>
</tr>
<tr>
<td>0.5107</td>
<td>antec:subj</td>
<td>0.6856</td>
</tr>
<tr>
<td>0.4760</td>
<td>conj:adversat-adverb</td>
<td>0.6239</td>
</tr>
<tr>
<td>0.4485</td>
<td>conj:adversat-subjun</td>
<td>0.4989</td>
</tr>
<tr>
<td>0.4120</td>
<td>antec:obj</td>
<td>0.4709</td>
</tr>
</tbody>
</table>
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


S. Evert, 2005. The CQP Query Language Tutorial. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, April. CWB version 2.2.b90.


References


Leech, G., M. Hundt, C. Mair, & N. Smith (2009). Change in Contemporary English. A


