Cross-linguistic Analysis of Cohesion
variation across production types and registers

Ekaterina Lapshinova-Koltunski and Kerstin Kunz

Saarland University, Heidelberg University
22 May 2013, Santiago de Compostela
Acknowledgement

Research Project

**GECCo: German-English Contrasts in Cohesion**
supported by the DFG

Project Team:
- Kerstin Kunz
- Ekaterina Lapshinova-Koltunski
- Marilisa Amoia
- Katrin Menzel
- Erich Steiner

FR 4.6 Applied Linguistics, Interpreting and Translation Studies

[www.gecco.uni-saarland.de](http://www.gecco.uni-saarland.de)
Aims and Motivation

Goal of Present Study

- subtypes or functions (cf. Kunz, 2009; Kunz and Steiner, 2012)
- across:
  - 1 languages: English vs. German
  - 2 registers: different text types
  - 3 production types: originals vs. translations
Goal of Present Study

**cohesive reference:**

- **types**: personal, demonstrative, comparative (cf. Halliday & Hasan, 1976)
- **subtypes or functions** (cf. Kunz, 2009; Kunz and Steiner, 2012)

**across:**

1. **languages**: English vs. German
2. **registers**: different text types
3. **production types**: originals vs. translations
Present Study: Linguistic variation

Hypotheses:

- variation is lower between GO vs GTRANS than EO vs GTRANS
- we expect more variation in form and function on the fine-grained level (cf. Kunz and Steiner, 2012).

Research Questions:

- Between which subcorpora are the greatest differences: across languages, registers or production types? languages or originals vs translations?
- Which features cause these differences?
- What is the most prominent difference between originals and translations?
- Are differences due to interference or rather to normalisation?
Corpus-based Analysis
Corpus-based Analysis

- Corpus Data
- Data Extraction
- Data Evaluation
## Data: GECCo Corpus

### subcorpora vs registers

<table>
<thead>
<tr>
<th>subcorpora</th>
<th>registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(imported from CroCo)</td>
<td>FICTION, ESSAY</td>
</tr>
<tr>
<td>EO 🇪🇸</td>
<td>INSTR, POPSCI</td>
</tr>
<tr>
<td>GO 🇩🇪</td>
<td>TOU, WEB</td>
</tr>
<tr>
<td>ETRANS 🇩🇪 → 🇬🇧</td>
<td>SHARE, SPEECH</td>
</tr>
<tr>
<td>GTRANS 🇪🇸 → 🇬🇧</td>
<td></td>
</tr>
<tr>
<td>(collected)</td>
<td></td>
</tr>
<tr>
<td>EO-SPOKEN 🇪🇸</td>
<td>INTERVIEW, ACADEMIC</td>
</tr>
<tr>
<td>GO-SPOKEN 🇩🇪</td>
<td>FORUM, TALKSHOW</td>
</tr>
</tbody>
</table>

### GECCo annotation levels

1) **word**: ⇒ *word, lemma, pos*

2) **chunk**: ⇒ *sentences, syntactic chunks, clauses, cohesive devices*

3) **text**: ⇒ *registers*

4) **extralinguistic**: ⇒ *register analysis, speaker information*
## Data: GECCo Corpus

### subcorpora

<table>
<thead>
<tr>
<th>subcorpora</th>
<th>registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(imported from CroCo)</td>
<td></td>
</tr>
<tr>
<td>EO 🇪🇸 🇳🇴</td>
<td>FICTION, ESSAY</td>
</tr>
<tr>
<td>GO 🇪🇸 🇩🇪</td>
<td>INSTR, POPSCI</td>
</tr>
<tr>
<td>ETRANS 🇪🇸 → 🇬🇧</td>
<td>TOU, WEB</td>
</tr>
<tr>
<td>GTRANS 🇪🇸 → 🇬🇧</td>
<td>SHARE, SPEECH</td>
</tr>
<tr>
<td>(collected)</td>
<td></td>
</tr>
<tr>
<td>EO-SPOKEN 🇪🇸 🇳🇴</td>
<td>INTERVIEW, ACADEMIC</td>
</tr>
<tr>
<td>GO-SPOKEN 🇪🇸 🇳🇴</td>
<td>FORUM, TALKSHOW</td>
</tr>
</tbody>
</table>

### GECCo annotation levels

1) **word**: ⇒ *word, lemma, pos*
2) **chunk**: ⇒ *sentences, syntactic chunks, clauses, cohesive devices*
3) **text**: ⇒ *registers*
4) **extralinguistic**: ⇒ *register analysis, speaker information*
Corpus Annotation: Reference

- **reference_type** – types of reference:
  - personal
  - demonstrative
  - comparative

- **reference_func** – functional subtypes of reference:
  - *it/es* (endophoric and exophoric)
  - head
  - modifier
  - local
  - temporal
  - pronominal adverb
  - general
  - particular
Corpus Extraction: Register Distribution

> group Last match reference_type by match text_register;

<table>
<thead>
<tr>
<th>Register</th>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>FICTION</td>
<td>pers</td>
<td>1376</td>
</tr>
<tr>
<td>POPSCI</td>
<td>pers</td>
<td>804</td>
</tr>
<tr>
<td>SPEECH</td>
<td>dem</td>
<td>791</td>
</tr>
<tr>
<td>POPSCI</td>
<td>dem</td>
<td>706</td>
</tr>
<tr>
<td>FICTION</td>
<td>dem</td>
<td>670</td>
</tr>
</tbody>
</table>

> group Last match reference_func by match text_register;

<table>
<thead>
<tr>
<th>Register</th>
<th>Function</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>FICTION</td>
<td>person-endophoric</td>
<td>1095</td>
</tr>
<tr>
<td></td>
<td>possessive-endophoric</td>
<td>613</td>
</tr>
<tr>
<td></td>
<td>it-endophoric</td>
<td>360</td>
</tr>
<tr>
<td>SPEECH</td>
<td>modifier</td>
<td>294</td>
</tr>
<tr>
<td>ESSAY</td>
<td>particular</td>
<td>261</td>
</tr>
<tr>
<td>POPSCI</td>
<td>modifier</td>
<td>259</td>
</tr>
<tr>
<td>SHARE</td>
<td>particular</td>
<td>255</td>
</tr>
<tr>
<td>POPSCI</td>
<td>particular</td>
<td>238</td>
</tr>
<tr>
<td>SHARE</td>
<td>possessive-endophoric</td>
<td>235</td>
</tr>
<tr>
<td>TOU</td>
<td>possessive-endophoric</td>
<td>230</td>
</tr>
</tbody>
</table>
Data Evaluation

Correspondance Analysis:

- **Input**: frequencies of cohesive devices across registers and production types
- **Output**: a two dimensional graph with:
  - **arrows** for the observed feature frequencies
  - **points** for registers across production types
- **Interpretation**:
  - the length of the **arrows** indicates how pronounced a particular feature is
  - the position of the **points** in relation to the **arrows** indicates the relative importance of a feature for a register.
  - the **arrows** pointing in the direction of an axis indicate a high contribution to the respective dimension

cf. (Glynn, 2012)
Analysen
Correspondence Analysis

EO vs GO vs ETRANS vs GTRANS
Correspondence Analysis

[Diagram showing a scatter plot with various labeled points and axes labeled.]
Correspondence Analysis

Observations for $x$-axis separation:

1. **EO/GO/ETRANS/GTRANS**: FICTION
   - **EO/GTRANS**: WEB
   - **EO**: SPEECH
   - **ETRANS**: POPSCI
   - shared features: pers. head, pers. modifier and *it*-exophoric
     - most prominent: pers. head

2. **EO/GO/ETRANS/GTRANS**: ESSAY, INSTR, SHARE, TOU
   - **EO/GO/GTRANS**: POPSCI
   - **GO/GTRANS/ETRANS**: SPEECH
   - **GO/ETRANS**: WEB
   - shared features: all dem. and comp.
     - most prominent: comp. particular
Correspondence Analysis

**Observations for y-axis separation:**

1. **GO/GTRANS:** ESSAY, FICTION, POPSCI, TOU  
   **GO:** INSTR, SHARE, SPEECH, WEB  
   shared features: pers. head, pers. modifier, dem. local, dem. pronadv, dem. temporal, comp. particular  
   most prominent: dem. pronadv and dem. local

2. **EO/ETRANS/GTRANS:** INSTR, SHARE, SPEECH, WEB  
   **EO/ETRANS:** ESSAY, FICTION, POPSCI, TOU  
   shared features: pers. *it*-endo/exophoric, dem. head, dem. modifier, comp. general  
   most prominent: comp. general

**both y and x-axis:** dem. modifier
Correspondence Analysis

Interpreting Results

- **x-axis:**
  - separation between different registers
  - translations show differences and similarities from/with originals in both languages
  - **most prominent features:** pers. head and comp. particular

- **y-axis:**
  - clear separation between English and German originals
  - English translations are similar to English originals ⇒ **normalisation**?
  - German translations show more variation:
    - some registers similar to English originals ⇒ **interference**?
    - some registers similar to German originals ⇒ **normalisation**?
  - **most prominent features:** dem. pronadv, dem. local and comp. general
Analyses

Young men on the roof tops changed their tune.
Analyses

Young men on the roof tops changed their tune.

Hier keimte das Motiv meiner Tat.

Nach der Hinrichtung ihres Mannes fand auch Frau Groscurr...
Probleme treten auf, wenn mehrere Betriebssysteme auf einem Rechner installiert wurden.
If you see **such** interference, move the monitors apart until the interference disappears.

Noch **gravierendere** Probleme treten auf, wenn mehrere Betriebssysteme auf einem Rechner installiert wurden.
Denn **hierin** sind wir uns in der deutschen "financial community" einig.
I welcome **this** opportunity to clarify for you. Professors told me their concerns about those detained by the Coalition. **It** is a familiar complaint.

Denn **hierin** sind wir uns in der deutschen "financial community" **einst"
Discussion and Conclusions

Discussion

Research Questions:
1. Between which subcorpora are the greatest differences?
2. Which features cause these differences?
3. What is the most prominent difference between originals and translations?
4. Are differences due to interference or rather to normalization?
Discussion

Discussion

Research Questions:

1. Between which *subcorpora* are the greatest differences?
2. Which *features* cause these differences?
3. What is the most *prominent difference* between originals and translations?
4. Are differences due to *interference* or rather to *normalisation*?
Discussion

Research Questions:

1. Between which subcorpora are the greatest differences: across languages, registers or production types?
   ⇒ greatest differences between original subcorpora! translations are in between but ETRANS is closer to EO

2. Which features cause these differences?
   ⇒ ENGLISH:
   preference for pers. reference and comp. general and dem. modifier
   ⇒ GERMAN:
   preference for dem. pron. adverbs + dem. adverbials and comp. particular
Research Questions:

3 What is the most prominent difference between originals vs. translations (of the same language)?

register-dependent:
- GTRANS-FICTION: more pers. heads and modifiers, less pron. adverbials and loc. dem. than GO
- GTRANS-SPEECH: more pers. modifiers, dem. modifiers, and es-exophoric than GO
- GTRANS INSTR: less temp. and loc. adverials and less comp. particular
Discussion

Research Questions:

4. Are differences due to interference or rather to normalisation? language-/translation direction-dependent:

- EO $\Rightarrow$ GTRANS:
  1. strong interference
  2. normalisation (=exaggeration of TL Conventions) for particular registers on the other hand
  3. lower distributions than both original subcorpora
     $\Rightarrow$ strongly depends on register and devices of reference

  $\Rightarrow$ more heterogeneity!

- GO $\Rightarrow$ ETRANS:
  1. interference but not too such a strong degree
  2. ETRANS generally shows more commonalities to EO

  $\Rightarrow$ less distinct properties of translation,
  less dependence on register
Thank you!

Questions? Comments? Suggestions?
Ekaterina Lapshinova
e.lapshinova@mx.uni-saarland.de
Kerstin Kunz
kerstin.kunz@iued.uni-heidelberg.de