Researching into Academic Writing "at home": the ChemCorpus, compilation and exploitation

Josef Schmied
English Language & Linguistics
Chemnitz University of Technology
http://www.tu-chemnitz.de/phil/english/ling/presentations Js.php
1. Concepts of and in Academic English

1.1 Academic English, an old and a new field

English for Academic Purposes (EAP), ESP, EIL, lingua franca, etc.

new theoretical basis: from Swales to Hyland

= initiation into the culture and practices of various ‘discourse communities’ (Swales 1990).

From a social constructionist point of view,
‘writing is always a personal and socio-cultural act of identity whereby writers both signal their membership in a range of communities, as well as express their own creative presence’ (Hyland 2006: 35).

3 levels of communities:
undergraduate
graduate
postdoc
1.2 Discourses in Academic English

- Research discourse
- Instructional discourse
- Popular discourse
- Student discourse
1.3 The Genre Approach

**research “output”**
- research article
  - book reviews
  - project proposals
- conference presentations

**science “journalism”**
- popular science articles
  - popular blogs (David Crystal)
  - popular science films (Horizon)
- popular science books

**teacher “talk”/e-learning**
- ppt presentations
  - lectures
  - student presentations
- textbooks
  - Wikis
  - www pages

**student "literacy”**
- fieldwork notes / essays
- Mag/BA thesis
- seminar presentations

“Novice Academic English”
1.4 Three Levels of Academic Writing / EAP

**Student/Novice EAP: "digesting" research**
incl. independent search for appropriate information, its critical evaluation and media-specific presentation
prototype genre: essay (1-5,000 words)

**Doctoral EAP: "developing" research**
national?
incl. a long-term independent research agenda and presentation of result of a sophisticated innovative project
with subject-specific conventions (MLA, ASA/APA)
prototype genre: PhD thesis (50,000-words?)

**Research EAP: participating in research discourse**
international
incl. peer-review procedures and the corresponding discourse with subject- and journal-specific conventions
prototype genre: research article (5,000-20,000 words)
2. Data and Examples: (sub-)genres of Academic English

2.1 Principle text genres?

**written** (excl. written to be spoken)
- assignments: homework/tasks in student portfolio
- seminar papers (10/20 pages; 5000 words)
- MagThesis (>50 pages; 25000 words)
- MagWritten (2000 words)
- Assignments: homework/tasks in student portfolio

**spoken** (incl. written to be spoken)
- placement test interviews
- ppt seminar presentations
- ppt presentation videos
- Mag/BA oral?
2.2. Database Chemnitz Corpus of Student English (ChemCorpus)

data collection in our own department (for BA theses and beyond)

written:
• MagTheses 35
• BAtheses 11
• MAtheses 3
• MagWritten 52

new translation and essay-writing tests on-line?

written to be spoken:
• ppt seminar presentation

oral:
• placement test interviews: ?
• ppt presentations: BA oral Presentation 15 min
### 2.2 Database ChemCorpus = Chemnitz Corpus of (written) Student English

<table>
<thead>
<tr>
<th>Corpus</th>
<th># files</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachelor Year 2 Term Papers Linguistics CUT</td>
<td>18</td>
<td>75,528</td>
</tr>
<tr>
<td>Bachelor Theses Linguistics CUT</td>
<td>11</td>
<td>143,692</td>
</tr>
<tr>
<td>Magister Exams Linguistics CUT</td>
<td>52</td>
<td>103,732</td>
</tr>
<tr>
<td>Magister Theses Linguistics CUT</td>
<td>24</td>
<td>652,869</td>
</tr>
<tr>
<td>Magister Theses Cultural Studies/Literature CUT</td>
<td>10</td>
<td>292,537</td>
</tr>
<tr>
<td><strong>total CUT</strong></td>
<td><strong>115</strong></td>
<td><strong>1,268,358</strong></td>
</tr>
</tbody>
</table>
3. Methodologies of Annotation and Analysis

3.1 Theoretical Approaches

functional approaches by Halliday

style pattern approaches from Kaplan to Biber

norms and critical discourses:
   from Quirk and Kachru to Fairclough and Phillipson

what do we compare with?
   EAP non-native speakers with native speakers?
   EAP users on the same level / with highest levels?
   EAP users with ideals?
3.2 Corpus-linguistic Methodologies

A clear analysis of the form–function relationship is the basis for corpus-linguistic studies since corpus-analysis tools (e.g., AntConc) are based on form.

But which forms matter?
- Functional: more persuasive/audience-adaptive (e.g., hedges)?
- Style: less complex (e.g., what)?

Formal vs. pragmatic features in argumentative styles?
## 3.3 Linguistic Features: function → form

<table>
<thead>
<tr>
<th>research articles</th>
<th>instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>• novelty</td>
<td>• macro-/micro-structuring</td>
</tr>
<tr>
<td></td>
<td>• meta-discourse</td>
</tr>
<tr>
<td></td>
<td>• interpersonal</td>
</tr>
<tr>
<td></td>
<td>• attitudinal</td>
</tr>
<tr>
<td></td>
<td>(compare MICASE, BASE)</td>
</tr>
</tbody>
</table>

### novelty
- in lit. review evaluation → stance, hedges, boosters
- modal aux/adv

### complexity
- in ontology lexical hypernyms/hyponyms

<table>
<thead>
<tr>
<th>science journalism</th>
<th>student papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>• metaphors</td>
<td>• argumentation structure</td>
</tr>
<tr>
<td>• multimedia with</td>
<td>• coherence →</td>
</tr>
<tr>
<td>text/image/sound</td>
<td>cohesion by pronoun</td>
</tr>
<tr>
<td>(??)</td>
<td>chains, clause adv.</td>
</tr>
<tr>
<td></td>
<td>(compare BAWE)</td>
</tr>
</tbody>
</table>
3.4 Research Question: Convergence or Diversity?

How much convergence do we want in research English?
How much diversity do we accept in novice English?
Which features do we want to reduce or increase?
4. Case study:  
4.1 Stance in ChemCorp

The expression of stance

**epistemic stance** (certainty (low, moderate and high level of certainty)  
adverbs: e.g. *maybe, certainly, definitely*  
verbs: e.g. *seem, indicate, show*  
adjectives: e.g. *evident, possible, obvious*  
nouns: e.g. *assumption, possibility, likelihood*  
modals: e.g. *might, may, could*

**attitudinal stance** (attitudes, feelings or value judgements)  
adverbs: e.g. *unfortunately, interestingly, happily*  
verbs: e.g. *hope, succeed, fail*  
adjectives: e.g. *appropriate, important, surprising*  
nouns: e.g. *triumph, intention, expectation*
## Stance in ChemCorp: Adverbs

<table>
<thead>
<tr>
<th>Adverb</th>
<th>MagTheses</th>
<th>MagWritten</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per 100000</td>
<td>Per 100000</td>
</tr>
<tr>
<td>possibly</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>maybe</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>perhaps</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>probably</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>definitely</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>certainly</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>of course</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>indeed</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>surely</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>clearly</td>
<td>50</td>
<td>19</td>
</tr>
<tr>
<td>obviously</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>apparently</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>181</strong></td>
<td><strong>102</strong></td>
</tr>
</tbody>
</table>
Stance in ChemCorp MagWritten

features of epistemic and attitudinal stance in MagWritten

As all these definitions are pretty vague I [presence-author] think [low to moderate certainty] that a better approach on understanding bilingualism can [possibility] be made by looking at the different dimensions. (S07_D_H)

For the purpose of discussing universals in linguistics, Greenberg’s approach is the appropriate one. (W0708_Y_J)

Interpreting these results I [presence-author] would also assume [low certainty], that a woman tends to [low certainty] use more affective tags in order to involve her readership. (S09_C_S)
4. Case study 2: Hedging in German Student Writing: female vs. male in MA theses?

Bayer, Dana (2009), *Gender and Hedging in Academic Writing*. BA Thesis Chemnitz, p. 32
### 4.3 Case study 3: concessive/contrastive markers in the ChemCorpus and comparable corpora

<table>
<thead>
<tr>
<th>(sub-)corpus</th>
<th>BAWE L1</th>
<th>BAWE non-L1</th>
<th>BAWE Ling.</th>
<th>MICUSP L1</th>
<th>MICUSP Ling.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BA term papers year 2</td>
<td>0.22</td>
<td>0.64</td>
<td>0.30</td>
<td>0.61</td>
<td>0.47</td>
</tr>
<tr>
<td>BA Theses Linguistics</td>
<td>0.33</td>
<td>0.81</td>
<td><strong>0.11</strong></td>
<td>0.73</td>
<td>0.53</td>
</tr>
<tr>
<td><em>Magister</em> written exams</td>
<td>0.24</td>
<td>0.77</td>
<td><strong>0.14</strong></td>
<td>0.64</td>
<td>0.46</td>
</tr>
<tr>
<td><em>Magister</em> Theses Linguistics</td>
<td>0.24</td>
<td>0.56</td>
<td>0.36</td>
<td>0.60</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Chi-square values CUT vs. Reference Corpora (cf. Wagner 2011: 35)
5. Conclusion for corpus-linguists

- some “accessible” language variables:
  - modal aux - epistemic usage?
  - prepositions - *of for NPs?
  - sentence adverbs - */y,

- some good socio-biographical/text variables:
  - gender
  - short-term: students through 5 years of study
  - long-term: students over the years
5. **Conclusion for students of English**

ChemCorpus: a good database, because
- students can combine theory and practice
- analysis raises awareness, increases editing skills
- allows comparisons with
  - grammar/statistics: *Longman Grammar of Written & Spoken English* (Biber et al. 1999)
  - corpora: BAWE, MICASE, BNC, ICLE
  - project partners in CZ, I, FR?
References


+ PhD theses