A corpus-based approach to operationalizing and assessing writing proficiency in the academic register:
The case of reporting verbs

Marcus Callies
Roadmap

1. Learner corpora in Language Testing and Assessment
2. Describing a data-driven approach
3. Case study on reporting verbs
4. Conclusion
1. Learner corpora in LTA (1)

- learner corpus: systematic collection of **authentic, continuous and contextualized language use** by foreign/second language learners, stored in electronic format (**uncontrolled, open-ended**)
- used in SLA research for almost two decades; **Learner Corpus Research (LCR)** = computer-aided approach to storage and processing of (mostly written) learner data
- enables collection and analysis of large amounts of data (more difficult in earlier SLA research that largely used experimental data) in order to ...
  - ... improve in-depth description of (advanced) interlanguages
  - ... give SLA theories a more solid empirical foundation (alongside with experimental data)
  - ... help produce tools and teaching materials designed for needs of specific learner populations
  - ... inform/complement assessment of L2 proficiency
1. Learner corpora in LTA (2)

- LCR responding to limitations of 'traditional' ways of assessing writing proficiency in LTA:
  - writing tasks (as part of tests)
  - expert raters
  - rating using scales
  - assigning CEFR-level

- subjectivity and variability of human rating vs. "objective", quantifiable linguistic descriptors

- recent research strand using learner corpora to inform, validate, and advance L2 proficiency assessment based on CEFR:
  - can-do statements do not provide language-specific, fine-grained linguistic details regarding learners' skills in certain registers
  - need to identify corpus-based, quantifiable linguistic descriptors ('criterial features') to add "grammatical and lexical details of English to CEFR's functional characterisation of the different levels" (Hawkins & Filipovic 2012: 5)
1. Learner corpora in LTA (3)

• three approaches:
  – corpus-informed
  – corpus-based
  – corpus-driven

• distinction based on three aspects:
  – how corpus data are actually put to use
  – aims and outcomes for LTA
  – degree of involvement of researcher in data retrieval, analysis and interpretation

→ no strict distinctions, may overlap and merge in some practises
### Corpus-informed approach

<table>
<thead>
<tr>
<th>use of data</th>
<th>aims &amp; outcomes for LTA</th>
<th>involvement of researcher</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus as reference source; provides practical information on learners' language use (= what they can do) at certain levels of proficiency</td>
<td>evidence to inform and validate test content and practices</td>
<td>high</td>
<td><em>Cambridge Learner Corpus</em> (CLC; Hawkey &amp; Barker 2004); <em>Pearson International Corpus of Academic English</em> (Ackermann, Biber &amp; Gray 2011)</td>
</tr>
</tbody>
</table>
## Corpus-based approach

<table>
<thead>
<tr>
<th>use of data</th>
<th>aims &amp; outcomes for LTA</th>
<th>involvement of researcher</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus as source of data for linguistic research, testing existing hypotheses about learner language</td>
<td>evidence used to identify set of distinct features or descriptors for differentiating proficiency levels (criterial features)</td>
<td>medium</td>
<td>CLC vs. <em>British National Corpus</em> (Hawkins &amp; Filipović 2012)</td>
</tr>
</tbody>
</table>
## Corpus-driven approach

<table>
<thead>
<tr>
<th>use of data</th>
<th>aims &amp; outcomes for LTA</th>
<th>involvement of researcher</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus driven = data driven (Francis 1993); no preconceptions/ hypotheses prior to corpus analysis; using computer techniques for data extraction and evaluation</td>
<td>evidence of proficiency based on statistical analyses largely independent of human rating</td>
<td>low</td>
<td><em>International Corpus of Learner English</em> (Wulff &amp; Gries 2011)</td>
</tr>
</tbody>
</table>
2. Describing a data-driven approach

Aims

- identify and operationalize fine-grained linguistic descriptors for assessment of writing proficiency in academic register
- data-driven approach to assessing proficiency partially independent of human rating
- "sophisticated language use in context" = implementing and operationalizing set of "positive linguistic properties" to determine what learners can do at advanced levels when writing for academic purposes
- combining three approaches in use of learner corpora for proficiency assessment: corpus-informed, corpus-based, and corpus-driven

Step 1

- select linguistic feature(s) that characterize academic prose (informed by corpus research on expert native-speaker usage)
- select descriptors in terms of keyness, operationalizability and if they remain problematic even for highly proficient L2 learners ("late acquired features")
- possible candidates for linguistic descriptors of academic writing:
  - specific constructions (verb-argument constructions, e.g. focus constructions, raising);
  - inanimate subjects (e.g. This paper discusses..., The results suggest that...)
  - phrases to express rhetorical functions (e.g. by contrast, to conclude, in sum)
  - reporting verbs (e.g. discuss, claim, suggest, argue etc.)
  - lexical co-occurrence patterns (e.g. conduct, carry out, undertake as typical verbal collocates of experiment, analysis, research)
Steps 2 & 3

- retrieve and analyse descriptors in academic learner writing (corpus-based)
- classify and assess written sample using statistical techniques (corpus-driven)
The corpus

- Language for Specific Purposes learner corpus containing discipline- and genre-specific texts; focus on written academic English ("academic learner writing")
- seven academic text types ("genres") produced as assignments in content courses by university students of English
- 50-100 texts per genre and L1 component
- L1 backgrounds so far: mostly German, Lithuanian, Russian, Polish, Turkish
lexico-grammatical choices to refer to writer (= author-agent) and others in *reporting events* in academic texts:

1) **We** will then analyze data specific to the case of Nicaraguan Sign Language to compare ...
2) **Neville and Buckingham** (1996) and later **Hyland** (1999) analyze some of these linguistic options for citation in detail ...
3) To understand why this failure occurred, **one** must consider two factors: the experimental design and ...
4) **This paper** will analyze the organizational structure of art-historical discourse through a study of seven texts about portrait paintings ...
5) Overall **it** could be argued that men have a higher social status than women due to using better forms of language.
6) Once a patient’s language abilities have been analyzed using linguistic assessment tests ...
Example: 'Agentivity' of academic writing (2)

- Biber & Conrad (2009: 162): almost no first person references in modern articles, but agentless passives and inanimate subjects common
- CALE: 18 single-authored linguistics term papers (62,300 words, approx. 2,940 sentences) written by German EFL learners
- compared with similar subset from *Michigan Corpus of Upper-Level Student Papers* (MICUSP; Römer & O'Donnell 2011)
- main findings:
  - significant underrepresentation of inanimate subjects in L2 writing (but preferred reporting strategies in L1 academic English)
  - overrepresentation of strategies to suppress agent (due to avoidance of 1. person pronouns), e.g. passive constructions with semantically 'empty' subject-placeholders:
    
    *There are two things to be discussed in this section.*
    
    *It can be observed that ...*
First person pronouns

![Bar chart showing first person pronoun usage](chart.png)
First person pronouns, subject placeholders and inanimate subjects

Callies (2013)
3. Case study: Reporting verbs in academic writing

- descriptor: lexical verbs frequently used to report facts and findings in academic writing = **reporting verbs** (aka research verbs, discourse verbs)
- crucial for reporting content, establishing other authors' and writer's own claims and situating these within published research = **high keyness**
- list of frequent reporting verbs drawn up from research literature; verbs extracted from corpus semi-automatically = **comparatively easy to operationalize**
- learners demonstrating high level of general language proficiency have limited inventory of reporting verbs in academic writing = **late-acquired**
From external criteria to linguistic descriptors

- **global measure**
  - pool of learners contributing texts to corpus; 
  - external measure: *institutional status*

- **local measures**
  - statistical techniques using linguistic descriptors

  descriptor: diversity of *reporting verbs*

  - diverse
    - (e.g. *say, state*)
  - diverse
    - (e.g. *claim, discuss, argue*, etc.)
3. Case study: Methodology

- **data:** 50 research papers produced by German EFL student writers at university; POS-tagged for corpus processing
- **list of 35 frequent reporting verbs in academic writing** compiled from academic word lists and research literature
- texts fed into AntConc, a corpus processing tool; target verbs extracted and counted (semi-)automatically
- texts ranked according to diversity of reporting verbs used:
  - size (number of tokens)
  - richness (number of types)
  - evenness (degree to which tokens are distributed equally across types)
- Simpson's Index of Diversity $D$: measure of diversity accounting for both richness and evenness; a figure between 0 and 1 = the greater the value, the greater the sample diversity
3. Case study: Results (all verb tokens)
3. Case study: Results (verb types per text)
3. Case study: Results (texts by $D$-score)
3. Case study: Results (types correlated with $D$)

Intermediate positive correlation between types and $D$ ($r=0.40$)
4. Conclusion

- learner corpora can inform, complement and possibly advance assessment of L2 proficiency vis-à-vis the CEFR
- identify and operationalize set of descriptors that are
  - language- and register-specific
  - quantifiable
  - subject to (semi-) automatic processing
- data-driven assessment of writing proficiency in academic register involving three steps:
  1. select linguistic features for academic prose in terms of keyness, operationalizability and "late acquisition" (*corpus-informed*)
  2. retrieve descriptors from corpus (*corpus-based*)
  3. classify and assess proficiency using statistical techniques (*corpus-driven*)
References


