John Benjamins Publishing Company

RESEARCH NOTE

The *Corpus of Academic Learner English* (CALE)

A new resource for the assessment of writing proficiency in the academic register

Marcus Callies and Ekaterina Zaytseva
University of Bremen

Learner corpora present an option to inform, supplement and advance the way language proficiency is operationalized and assessed, and may also be used in data-driven approaches to the assessment of writing proficiency that are largely independent of human rating. The aim of this contribution is twofold: first, to introduce a new Language-for-Specific-Purposes learner corpus, the *Corpus of Academic Learner English* (CALE), currently being compiled for the study of academic learner writing; and second, to illustrate how the CALE is useful in a text-centered, corpus-driven approach to the assessment of academic writing to achieve a higher degree of reliability in assessing language proficiency.

Keywords: learner corpus research, English for academic purposes, academic writing, language assessment, advanced language proficiency, criterial features, corpus-driven

1. The CALE: Design, composition and annotation

Many existing and widely-used learner corpora, such as the *International Corpus of Learner English* (ICLE; Granger, Dagneaux, Meunier, & Paquot, 2009), are general-purpose corpora in that they include learner texts of a general argumentative, creative or literary nature. The large majority of these texts do not represent academic writing in a narrow sense as they differ from academic prose in some important aspects. First, they are often loosely characterized as 'essays', a cover term for a general text type that is open to subjective interpretation (student writers may differ considerably in what they consider an essay) which makes a comparison to more specific academic text types difficult. Second, they are argumentative

Dutch Journal of Applied Linguistics 2:1 (2013), **126–132.** DOI 10.1075/dujal.2.1.11cal
ISSN 2211–7245 / E-ISSN 2211–7253 © John Benjamins Publishing Company

texts whose communicative purpose is not to inform but rather to argue for a certain position, to voice a personal opinion or to persuade an (unspecified) audience (see the list of the most popular topics given to the students who wrote texts for the ICLE; Granger et al., 2009, pp. 6 ff.). Third, several characteristic linguistic features that predominantly occur in academic prose (see Section 3 below) are either absent or rare in general-purpose learner corpora.

The CALE is a specialised learner corpus comprising discipline- and genre-specific texts and may therefore be considered a 'Language-for-Specific-Purposes learner corpus' (Granger & Paquot, 2013). It includes seven academic text types ('genres') produced by learners of English as a foreign language (EFL) in university courses (i.e. English linguistics, literary and cultural studies), see Figure 1. Corpora that contain comparable native speaker (NS) writing and may thus serve as control corpora for the CALE are the *Michigan Corpus of Upper-Level Student Papers* (MICUSP; Römer & O'Donnell, 2011; O'Donnell & Römer, 2012) and the corpus of *British Academic Written English* (BAWE; Alsop & Nesi, 2009).

Figure 1. Academic text types in the CALE

Currently, we are mostly collecting texts and bio data from German EFL students, but the corpus will be expanded to include data from EFL learners of other mother tongue (L1) backgrounds to enable cross-linguistic and typological comparisons.

The text classification developed for the CALE is comparable with the NS control corpora, but has clear(er) textual profiles, adopting the situational characteristics and linguistic features identified for academic prose by Biber and Conrad (2009). A text's communicative purpose/goal serves as the main classifying principle, which helps to set apart the seven genres in terms of (a) the text's general purpose, (b) its specific purpose(s), (c) the skills the author demonstrates, and (d) the author's stance. In addition, each text type is described in terms of (a) structural features, (b) length, and (c) functional features. Table 1 illustrates the profile for the genre 'abstract'.

Students submit their texts in electronic form (typically in MS-Word or PDF-format). Thus, some manual pre-processing of these incoming files is necessary.

Table 1. Profile for the text type 'abstract' in the CALE

communicative goal/purpose	features
a. *general purpose* informational — to inform	a. *structural* not structured into sections; appears at beginning of text it comes with; may also occur as stand-alone entity instead of full paper
b. *specific purpose(s)* captures essence of published research (why, how, what: research focus, methodology results/findings, conclusion & recommendations); should help reader to quickly ascertain purpose, content and usefulness of publication	b. *length* rather short (approx. 100–250 words), rarely exceeding 500 words
c. *skills* author demonstrates ability to extract and provide essential information in exhaustive and compelling way	c. *functional* self-contained piece of writing, can be understood independently from accompanying publication
d. *stance* author's opinion/evaluation absent	

Extensive 'non-linguistic' information (such as table of contents, list of references, tables, figures) is deleted and substituted by placeholder tags around their headings or captions. The body of the text is then annotated for meta-textual, i.e. underlying structural features (section titles, paragraphs, quotations, examples) with the help of annotation tools and annotation software like the UAM Corpus Tool.[1] The texts are also annotated for metadata, i.e. learner variables such as L1, age, gender, etc. which are collected through a written questionnaire. Each file also includes metadata that pertain to each individual text such as genre, type of course and discipline the text was written in, the setting in which the text was produced etc. This information is also collected with the help of a questionnaire. In addition, based on a function-to-form approach to the analysis of learner language, parts of the corpus will be annotated for linguistic features, e.g. rhetorical functions and the lexico-grammatical means to express them (see Section 3 below and Zaytseva (2011) for a discussion of the advantages of this approach).

[1] Available from http://www.wagsoft.com/CorpusTool/index.html

2. Using the CALE in a text-centered, corpus-driven approach to assess academic writing proficiency

The *Common European Framework of Reference for Languages* (CEFR), though highly influential in language testing and assessment, has recently been criticized for the way it defines proficiency levels using "can-do-statements" (see Callies, Zaytseva, & Present-Thomas, this volume). There is an increasing awareness among researchers of the need to add language-specific lexical and grammatical details to the functional characterisations of the proficiency levels of the CEFR. The aim is to identify more explicit and 'hard' linguistic descriptors or "criterial features" (Hawkins & Filipović, 2012) in order to make it possible to differentiate between proficiency levels as regards individual languages and learners' skills in specific registers. Among possible candidates for such features are e.g. different clause types and verbal complementation patterns (Hawkins & Filipović, 2012), the proficient use of constructions in their lexico-grammatical association patterns measured by means of collostructional analysis (Wulff & Gries, 2011), or the use of specific lexical verbs as reporting strategies in academic discourse as analysed by Callies (2013).

In line with this approach, we suggest a corpus-based implementation of several well-known characteristics of academic prose followed by a corpus-driven assessment of writing proficiency. Our basis is the construct "sophisticated language use in context" (Ortega & Byrnes, 2008) as a way to operationalize advancedness. Wulff and Gries's (2011) definition of accuracy as "proficient selection of constructions in their preferred constructional context in a particular target genre" (p. 61) would be a specific instance of sophisticated language use in context in which native-like proficiency is seen as a "gradual, probabilistic phenomenon that transcends a native-nonnative speaker divide" (p. 61).

The procedure of operationalizing linguistic descriptors in the CALE proceeds in a number of steps. The first step is to draw up a list of characteristic features of academic prose from which linguistic descriptors are selected in terms of their 'keyness' (i.e. how important and characteristic they are of the register) and their operationalizability (i.e. how well they can be retrieved from the corpora and subjected to statistical analysis). This step is based on a review of the pertinent research literature on academic writing, and studies that have identified some of the crucial features that remain problematic even for highly proficient L2 learners (similar to what Ortega and Byrnes (2008) call "late acquired features"). Some possible candidates are:

- specific constructions (verb-argument constructions, e.g. causative constructions, focus constructions, raising) (Callies, 2009; Gilquin, 2012; Wulff & Gries, 2011; Hawkins & Filipović, 2012);

- inanimate subjects (e.g. *This paper discusses…*, *The results suggest that…*) (Callies, 2013; Dorgeloh & Wanner, 2009; Master, 1991);
- phrases to express rhetorical functions (e.g. *by contrast*, *to conclude*, *in sum*) (Paquot, 2010);
- so-called "reporting verbs" (e.g. *discuss*, *claim*, *suggest*, *argue* etc.) (Callies, 2013; Granger & Paquot, 2009);
- lexical co-occurrence patterns (e.g. *conduct*, *carry out*, *undertake* as typical verbal collocates of *experiment*, *analysis*, *research*) (Ackermann, Biber, & Gray, 2011; Durrant, 2009).

Second, a feature is retrieved semi-automatically from the corpus and subjected to statistical analyses to identify clusters formed by samples of data (i.e. learner texts) that demonstrate the highest degree of similarity as to the occurrence of one or more features. Each feature (or descriptor) serves as a local measure of proficiency (see Figure 2). For example, in the case of reporting verbs the descriptor takes into account the diversity of the verbs used by a writer, i.e. how many different verbs are used and how often these occur. Depending on how diverse the use of reporting verbs is in comparison to the other texts in the corpus, individual papers will be placed into one or more clusters on the basis of the data subjected

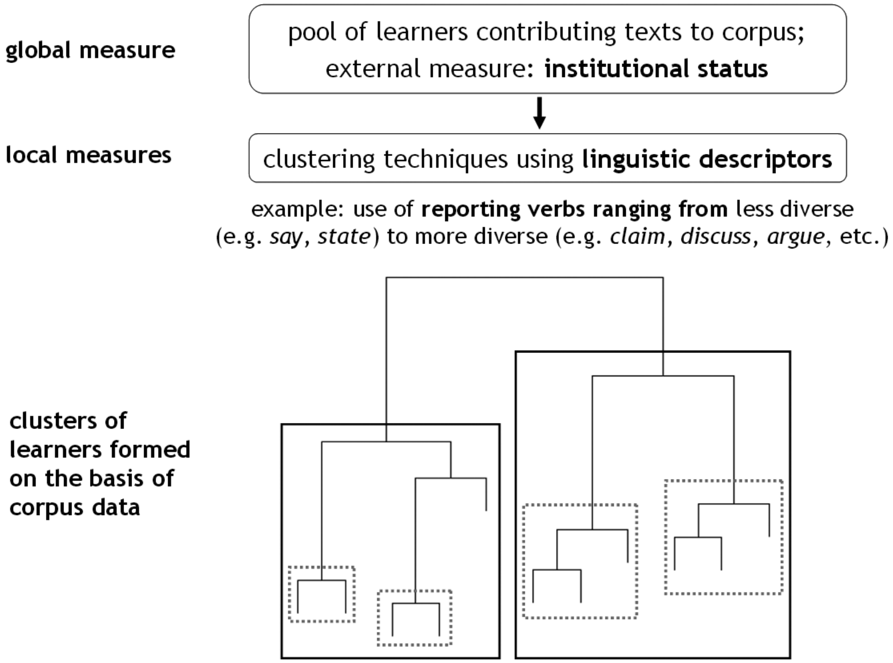
global measure
pool of learners contributing texts to corpus; external measure: institutional status
local measures
clustering techniques using linguistic descriptors
example: use of reporting verbs ranging from less diverse (e.g. say, state) to more diverse (e.g. claim, discuss, argue, etc.)
clusters of learners formed on the basis of corpus data

Figure 2. Applying linguistic descriptors and clustering techniques in the assessment of advanced writing proficiency

to the cluster analyses as shown in Figure 2. As a result, the clusters that have been formed represent gradual usage-based information on advancedness as anchored in academic writing.

Learners whose language use is significantly different from that of the rest will be identified by small, separate clusters or 'outliers'. As a result of multiple analyses involving a variety of descriptors, learner performance can be then visualized as a continuum of advancedness, where an individual learner will be assigned a particular level of general language proficiency (e.g. from (high) intermediate to near-native). Depending on the purposes of language testers or SLA researchers, this kind of corpus analysis can be extended to comparable novice and expert NS writing. Information gained in the course of such a corpus-driven approach can subsequently be used to specify the description of the advanced level within the CEFR and besides, to inform

- various stages of testing such as test development by providing evidence of what to measure at this level of proficiency in academic writing;
- test design by helping to develop realistic tasks;
- rating/evaluation by providing usage-based, empirical information;
- the interpretation of results and assessment of a proficiency level.

References

Ackermann, K., Biber, D., & Gray, B. (2011). *An academic collocation list*. Paper presented at Corpus Linguistics 2011, Birmingham, UK.

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, *4*(1), 71–83.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Callies, M. (2009). *Information highlighting in advanced learner English. The syntax-pragmatics interface in second language acquisition*. Amsterdam: John Benjamins.

Callies, M. (2013). Agentivity as a determinant of lexico-syntactic variation in L2 academic writing. *International Journal of Corpus Linguistics, 18*(3).

Callies, M., Zaytseva, E., & Present-Thomas, R. (this volume). Writing assessment in higher education: Making the framework work.

Dorgeloh, H., & Wanner, A. (2009). Formulaic argumentation in scientific discourse. In R. Corrigan, E.A. Moravcsik, H. Ouali, & K.M. Wheatley (Eds.), *Formulaic language: Volume 2. Acquisition, loss, psychological reality, and functional explanations* (pp. 523–544). Amsterdam: John Benjamins.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169.

Gilquin, G. (2012). Lexical infelicity in causative constructions. Comparing native and learner collostructions. In J. Leino, & R. von Waldenfels (Eds.), *Analytical causatives* (pp. 41–64). München: Lincom.

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing. At the interface of corpus and discourse* (pp. 193–214). London: Continuum.

Granger, S., & Paquot, M. (2013). Language for specific purposes learner corpora. In T.A. Upton, & U. Connor (Eds.), *Language for specific purposes. The encyclopedia of applied linguistics* (pp. 3142–3146). New York: Blackwell.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Hawkins, J.A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.

Master, P. (1991). Active verbs with inanimate subjects in scientific prose. *English for Specific Purposes*, *10*(1), 15–33.

O'Donnell, M., & Römer, U. (2012). From student hard drive to web corpus (Part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *7*(1), 1–18.

Ortega, L., & Byrnes, H. (2008). The longitudinal study of advanced L2 capacities: An introduction. In L. Ortega, & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 3–20). New York: Routledge.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.

Römer, U., & O'Donnell, M. (2011). From student hard drive to web corpus (Part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*(2), 159–177.

Wulff, S., & Gries, S. (2011). Corpus-driven methods for assessing accuracy in learner production. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 61–87). Amsterdam: John Benjamins.

Zaytseva, E. (2011). Register, genre, rhetorical functions: Variation in English native-speaker and learner writing. In H. Hedeland, T. Schmidt, & K. Wörner (Eds.), *Multilingual resources and multilingual applications* (pp. 239–242). Hamburg: University of Hamburg.