The *Corpus of Academic Learner English* (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties

Marcus Callies, Ekaterina Zaytseva
Johannes-Gutenberg-Universität Mainz, Department of English and Linguistics
Jakob-Welder-Weg 18, 55099 Mainz
E-mail: mcallies@uni-mainz.de, zaytseve@uni-manz.de

Abstract

This paper introduces the *Corpus of Academic Learner English* (CALE), a Language for Specific Purposes learner corpus that is currently being compiled for the quantitative and qualitative study of lexico-grammatical variation patterns in advanced learners' written academic English. CALE is designed to comprise seven academic genres produced by learners of English as a foreign language in a university setting and thus contains discipline- and genre-specific texts. The corpus will serve as an empirical basis to produce detailed case studies that examine individual (or the interplay of several) determinants of lexico-grammatical variation, e.g. semantic, structural, discourse-motivated and processing-related ones, but also those that are potentially more specific to the acquisition of L2 academic writing such as task setting, genre and writing proficiency. Another major goal is to develop a set of linguistic criteria for the assessment of advanced proficiency conceived of as "sophisticated language use in context". The research findings will be applied to teaching English for Academic Purposes by creating a web-based reference tool that will give students access to typical collocational patterns and recurring phrases used to express rhetorical functions in academic writing.

Keywords: learner English, academic writing, lexico-grammatical variation, advanced proficiency

1. Introduction

Recently, second language acquisition (SLA) research has seen an increasing interest in advanced stages of acquisition and questions of near-native competence. Corpus-based research into learner language (Learner Corpus Research, LCR) has contributed to a much clearer picture of advanced interlanguages, providing evidence that learners of various native language (L1) backgrounds have similar problems and face similar challenges on their way to near-native proficiency. Despite the growing interest in advanced proficiency, the fields of SLA and LCR are still struggling with i) a definition and clarification of the concept of "advancedness", ii) an in-depth description of ALVs, and iii) the operationalization of such a description in terms of criteria for the assessment of advancedness. In this paper, we introduce the *Corpus of Academic Learner English* (CALE), a Language for Specific Purposes learner corpus that is currently being compiled for the quantitative and qualitative study of lexico-grammatical variation patterns in advanced learners' written academic English.

2. Corpus design and composition

Already existing learner corpora, such as the *International Corpus of Learner English* (Granger et al., 2009) include learner writing of a general argumentative, creative or literary nature, and thus not academic writing in a narrow sense. Thus, several patterns of variation that predominantly occur in academic prose (or are subject to the characteristic features of this register) are not represented at all or not frequently enough in general learner corpora. CALE is designed to comprise academic texts produced by learners of English as a foreign language (EFL) in a university setting. CALE may therefore be considered a Language for Specific Purposes learner corpus, containing discipline- and genre-specific texts (Granger & Paquot, forthcoming). Similar corpora that contain native speaker (NS) writing and may thus serve as control corpora for CALE are the *Michigan Corpus of Upper-Level Student Papers* (MICUSP, Römer & Brook O'Donnell, forthcoming) and the *British Academic Written English corpus* (BAWE, Alsop & Nesi, 2009).

CALE's seven academic text types ("genres") are written as assignments by EFL learners in university courses, see Figure 1.

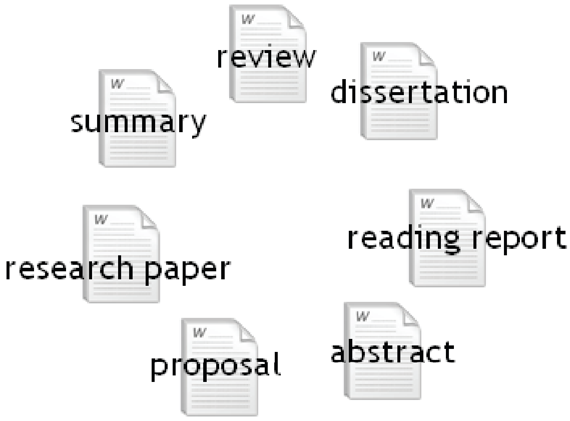

Figure 1: Academic text types in CALE

We are currently collecting texts and bio data from German, Chinese and Portuguese students, and are planning to include data from EFL learners of other L1 backgrounds to be able to draw cross-linguistic and typological comparisons as to potential L1 influence.

The text classification we have developed for CALE is comparable with the NS control corpora mentioned above, but we have created 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 or goal serves as the main classifying principle, which helps to set apart the seven genres in terms of

a) text's general purpose
b) its specific purpose(s)
c) the skills the author demonstrates, and
d) the author's stance.

In addition, we list the major features of each text type as to

a) structural features
b) length, and
c) functional features.

3. Corpus annotation

Students submit their texts in electronic form (typically in .doc, .docx or .pdf file format). Thus, some manual pre-processing of these incoming files is necessary. Extensive "non-linguistic" information (such as table of contents, list of references, tables and figures, etc.) 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, etc.) with the help of annotation tools. The texts are also annotated (in a file header) for metadata, i.e. learner variables such as L1, age, gender, etc. which are collected through a written questionnaire. The file header 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 that accompanies each text submitted to the corpus. In the future, we also intend to implement further linguistic levels of annotation, e.g. for rhetorical function or sentence type.

4. Research program

In the following sections, we outline our research program. We adopt a variationist perspective on SLA, combining a learner corpus approach with research on interlanguage variation and near-native competence.

4.1. The study of variation in SLA research

Interlanguages (ILs) as varieties in their own right are characterized by variability even more than native languages. Research on IL-variation since the late 1970s has typically focused on beginning and intermediate learners and on variational patterns in pronunciation and morphosyntax, i.e. the (un-)successful learning of actually invariant linguistic forms and the occurrence of alternations between native and non-native equivalent forms. Such studies revealed developmental patterns, interpreted as indicators of learners' stages of acquisition, and produced evidence that IL-variation co-varies with linguistic, social/situational and psycholinguistic context, and is also subject to a variety of other factors like individual learner characteristics and biographical variables (e.g. form and length of exposure to the L2).

Since the early 2000s there has been an increasing interest in issues of sociolinguistic and sociopragmatic variation in advanced L2 learners (frequently referred to as sociolinguistic competence), e.g. learners' use of dialectal forms or pragmatic markers (mostly in L2 French, see e.g. Mougeon & Dewaele, 2004; Regan, Howard & Lemée, 2009). This has marked both a shift from the study of beginning and intermediate to advanced learners, and a shift from the study of norm-violations to the investigation of differential knowledge as evidence of conscious awareness of (socio-)linguistic variation.

4.2. Advanced Learner Varieties (ALVs)

There is evidence that advanced learners of various language backgrounds have similar problems and face similar challenges on their way to near-native proficiency. In view of these assumed similarities, some of which will be discussed in the following, we conceive of the interlanguage of these learners as Advanced Learner Varieties (ALVs).

In a recent overview of the field, Granger (2008:269) defines advanced (written) interlanguage as "the result of a highly complex interplay of factors: developmental, teaching-induced and transfer-related, some shared by several learner populations, others more specific". According to her, typical features of ALVs are overuse of high frequency vocabulary and a limited number of prefabs, a much higher degree of personal involvement, as well as stylistic deficiencies, "often characterized by an overly spoken style or a somewhat puzzling mixture of formal and informal markers".

Moreover, advanced learners typically struggle with the acquisition of optional and/or highly L2-specific linguistic phenomena, often located at interfaces of linguistic subfields (e.g. syntax-semantics, syntax-pragmatics, see e.g. DeKeyser, 2005:7ff). As to academic writing, many of their observed difficulties are caused by a lack of understanding of the conventions of academic writing, or a lack of practice, but are not necessarily a result of interference from L1 academic conventions (McCrostie, 2008:112).

4.3. Patterns and determinants of variation in L2 academic writing

Our research program involves the study of L2 learners' acquisition of the influence of several factors on constituent order and the choice of constructional variants (e.g. genitive and dative alternation, verb-particle placement, focus constructions). One reason for this is that such variation is often located at the interfaces of linguistic subsystems, an area where advanced learners still face difficulties. Moreover, grammatical variation in L2 has not been well researched to date and is only beginning to attract researchers' attention (Callies, 2008, 2009; Callies & Szczesniak, 2008).

There are a number of semantic, structural, discourse-motivated and processing-related determinants that influence lexico-grammatical variation whose interplay and influence on speakers' and writers' constructional choices has been widely studied in corpus-based research on L1 English. Generally speaking, in L2 English these determinants play together with several IL-specific ones such as mother tongue (L1) and proficiency level, and in (academic) writing, some further task-specific factors like imagined audience (the people to whom the text is addressed), setting, and genre add to this complex interplay of factors, see Figure 2.

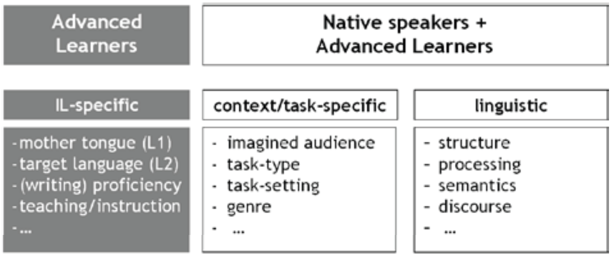

Figure 2: Determinants of variation in L1 and L2 academic writing

It is important to note at this point that differences between texts produced by L1 and L2 writers that are often attributed to the influence of the learners' L1 may in fact turn out to result from differences in task-setting (prompt, timing, access to reference works, see Ädel, 2008), and possibly task-instruction and imagined audience (see Ädel, 2006:201ff for a discussion of corpus comparability). Similarly, research findings as to learners' use of features that are more typical of speech than of academic prose have been interpreted as unawareness of register differences, but there is some evidence that the occurrence of such forms may also be caused by the influence of factors like the development of writing proficiency over time (novice writers vs. experts, see Gilquin & Paquot, 2008; Wulff & Römer, 2009), task-setting and -instruction, imagined audience and register/genre (e.g. academic vs. argumentative writing, see Zaytseva, 2011).

4.4. Case study

In this section, we provide an example of how lexico-grammatical variation plays out in L2 academic writing. In a CALE pilot study of the (non-) representation of authorship in research papers written by advanced German EFL learners, Callies (2010) examined agentivity as a determinant of lexico-grammatical variation in academic prose. He hypothesized that even advanced students were insecure about the representation of authorship due to a mixture of several reasons: conflicting advice by teachers, textbooks and style guides, the diverse conventions of different academic disciplines, students' relative unfamiliarity with academic text types and lack of linguistic resources to report events and findings without mentioning an agent. Interestingly, the study found both an overrepresentation of the first person pronouns *I* and *we*, but also an overrepresentation of the highly impersonal subject-placeholders *it* and *there* (often used in the passive voice) as default strategies to suppress the agent, see examples (1) and (2).

(1) There are two things to be discussed in this section.
(2) It has been shown that…

While this finding seems to be contradictory, it can be explained by a third major finding, namely the significant underrepresentation of inanimate subjects which are, according to Biber and Conrad (2009:162), preferred reporting strategies in L1 academic English, exemplified in (3) and (4).

(3) This paper discusses…
(4) Table 5 shows that…

Callies (2010) concluded that L2 writers have a narrower inventory of linguistic resources to report events and findings without an overt agent, and their insecurity and unfamiliarity with academic texts adds to the observed imbalanced clustering of first person pronouns, dummy-subjects and passives. The findings of this study also suggest that previous studies that frequently explain observed overrepresentations of informal, speech-like features by pointing to learners' higher degree of subjectivity and personal involvement (Granger, 2008) or unawareness of register differences (Gilquin & Paquot, 2008), may need to be supplemented by studies taking into account a more complex interplay of factors that also includes the limited choice of alternative strategies available to L2 writers.

5. Implications for language teaching and assessment

The project we have outlined in this paper has some major implications for EFL teaching and assessment. The research findings will be used to provide recommendations for EFL teachers and learners by developing materials for teaching units in practical language courses on academic writing and English for Academic Purposes. In the long run, we plan to create a web-based reference tool that will help students look up typical collocations and recurring phrases used to express rhetorical moves/functions in academic writing (e.g. giving examples, expressing contrast, drawing conclusions etc.). This application will be geared towards students' needs and can be used as a self-study reference tool at all stages of writing an academic text. Users will be able to access information in two ways: 1) form-to-function, i.e. looking up words and phrases in an alphabetical index to see how they can express rhetorical functions, and 2) function-to-form, i.e. accessing a list of rhetorical functions to find words and phrases that are typically used to encode them.

Most importantly, the tool will present in a comparative manner structures that emerged as problematic in advanced learners' writing, for example untypical lexical co-occurrence patterns and over- or underrepresented words and phrases, side by side with those structures that typically occur in expert academic writing. This will include information on the immediate and wider context of use of single items and multi-word-units.

While the outcome is thus particularly relevant for future teachers of English, it may also be useful for students and academics in other disciplines who have to write and publish in English. Unlike in the Anglo-American education system, German secondary schools and universities do not usually provide courses in academic writing in the students' mother tongue, so that first-year students have basically no training in academic writing at all.

It has been mentioned earlier that the operationalization of a quantitatively and qualitatively well-founded description of advanced proficiency in terms of criteria for the assessment of advancedness is still lacking. Thus, a major aim of the project is to develop a set of linguistic descriptors for the assessment of advanced proficiency. The descriptors and can-do-statements of the Common European Framework of Reference (CEFR) often appear too global and general to be of practical value for language assessment in general, and for describing advanced learners' competence as to academic writing in particular. Ortega and Byrnes (2008) discuss four ways in which advancedness has commonly been operationalised, ultimately favouring what they call "sophisticated language use in context", a construct that includes e.g. the choice among registers, repertoires and voice. This concept can serve as a basis for the development of linguistic descriptors that are characteristic of academic prose, e.g. the use of syntactic structures like inanimate subjects, phrases to express rhetorical functions (e.g. *by contrast, to conclude, in fact*), reporting verbs (*discuss, claim, suggest, argue, propose* etc.), and lexical co-occurrence patterns (e.g. *conduct*, *carry out* and *undertake* as typical verbal collocates of *experiment, analysis* and *research*).

6. References

Ädel, A. (2006): Metadiscourse in L1 and L2 English. Amsterdam: Benjamins.

Ädel, A. (2008): Involvement features in writing: do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M.B. Diez-Bedmar (Eds.), Linking up Contrastive and Learner Corpus Research. Amsterdam: Rodopi, pp. 35-53.

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

Biber, D., S. Conrad (2009): Register, Genre, and Style. Cambridge: Cambridge University Press.

Callies, M. (2008): Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In G. Gilquin, S. Papp & M.B. Diez-Bedmar (Eds.), Linking up Contrastive and Learner Corpus Research. Amsterdam: Rodopi, pp. 201-226.

Callies, M. (2009): Information Highlighting in Advanced Learner English. Amsterdam: Benjamins.

Callies, M. (2010): The (non-)representation of authorship in L2 academic writing. Paper presented at ICAME 31 "Corpus Linguistics and Variation in English", 26-30 May 2010, Giessen/Germany.

Callies, M., Szczesniak, K. (2008): Argument realization, information status and syntactic weight - A learner-corpus study of the dative alternation. In P. Grommes & M. Walter (Eds.), Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Tübingen: Niemeyer, pp. 165-187.

DeKeyser, R. (2005): What makes learning second language grammar difficult? A review of issues. Language Learning, 55(s1), pp. 1-25.

Gilquin, G., Paquot, M. (2008): Too chatty: Learner academic writing and register variation. English Text Construction, 1(1), pp. 41-61.

Granger, S. (2008): Learner corpora. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics. An international handbook, Vol. 1. Berlin & New York: Mouton de Gruyter, pp. 259-275.

Granger, S., Paquot, M. (forthcoming): Language for Specific Purposes learner corpora. In T.A. Upton & U. Connor (Eds.), Language for Specific Purposes. The Encyclopedia of Applied Linguistics. 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.

McCrostie, J. (2008): Writer visibility in EFL learner academic writing: A corpus-based study. ICAME Journal, 32, pp. 97-114.

Mougeon, R., Dewaele, J.-M. (2004): Patterns of variation in the interlanguage of advanced second language learners. Special issue of International Review of Applied Linguistics in Language Teaching (IRAL), 42(4).

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. New York: Routledge/Taylor & Francis, pp. 3-20.

Regan, V., Howard, M., Lemée, I. (2009): The Acquisition of Sociolinguistic Competence in a Study Abroad Context. Clevedon: Multilingual Matters.

Römer, U., Brook O'Donnell, M. (forthcoming): From student hard drive to web corpus: The design, compilation, annotation and online distribution of MICUSP. Corpora.

Wulff, S., Römer, U. (2009): Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. Corpora, 4(2), pp. 115-133.

Zaytseva, E. (2011): Register, genre, rhetorical functions: Variation in English native-speaker and learner writing. Hamburg Working Paper in Multilingualism.