Writing assessment in higher education

Making the framework work

Marcus Callies*, Ekaterina Zaytseva* and
Rebecca Present-Thomas**
*University of Bremen / **VU University Amsterdam

The importance of appropriate assessment methods for academic writing skills in higher education has received increasing attention in SLA research in recent years. Despite this, there is still relatively little understanding of how academic writing skills develop at the most advanced levels of proficiency. Use of the *Common European Framework of Reference for Languages* (CEFR) is one way to ensure the comparability of findings across research efforts and continue to move the field forward. This paper presents some key concepts and definitions from the fields of SLA and advancedness research, language assessment and corpus linguistics and introduces several papers that address writing assessment within the context of higher education.

Keywords: academic writing, SLA, assessment, CEFR, corpus linguistics

1. Introduction

In the 12 years since the *Common European Framework of Reference for Languages* (CEFR; Council of Europe, 2001) was published, it has been widely adopted in local and global contexts. Significant appeal of the CEFR is due to its multi-purpose nature, which allows the proficiency framework to remain relevant across users, levels and even languages. This flexibility is also, arguably, its largest pitfall when it comes to consistent use across diverse contexts. Though widely adopted across a range of settings, many have expressed concerns that the CEFR levels and descriptors are not suitable for direct application to any particular context (Alderson, 2007; Fulcher, 2004; Hulstijn, 2007). Arguably, it is flexible enough to be applied across a broad range of situations, yet too abstract to apply comprehensively to any one in particular.

Dutch Journal of Applied Linguistics 2:1 (2013), **1–15**. DOI 10.1075/dujal.2.1.03cal
ISSN 2211–7245 / E-ISSN 2211–7253

One such context in which the CEFR, as published, fails to provide thoroughly descriptive and diagnostic information about user proficiency is the construct of academic writing. This paper first serves to define this construct with respect to the concept of 'advancedness' inherent to academia. Next, we will provide a brief, critical overview of how this construct has been measured in previous studies and the validity of these measurements. Finally, we will outline how the CEFR has so far been incorporated into research on academic writing through the linking of existing examinations to the CEFR and the use of the CEFR levels in corpus-linguistic analyses.

2. Academic writing and advanced language proficiency

Academic writing/prose can generally be conceived of as "any writing that fulfills a purpose of education in a college or university (…); writing in response to an academic assignment, or professional writing that trained 'academics' — teachers and researchers — do for publications read and conferences attended by other academics" (Thaiss & Zawacki, 2006, p. 4). In addition, the following features are characteristic of this register (Department of Translation Studies, University of Tampere, Finland, n.d.):

- it represents structured research written by scholars for other scholars (with all university writers being scholars in this context)
- it addresses topic-based research questions of interest to anyone who is seeking factually-based, objectively-presented information on a particular topic
- its objective is the creation of new knowledge via (a) a review of what is currently known about a given topic as (b) the foundation for the author's new views or perspectives on the topic.

Biber, Johansson, Leech, Conrad and Finegan (1999) and Biber and Conrad (2009) consider academic prose "a very general *register*, characterized as written language that has been carefully produced and edited, addressed to a large number of readers who are separated in time and space from the author, and with the primary communicative purpose of presenting information about some topic" (Biber & Conrad, 2009, p. 32; our emphasis). "Register" is used as a cover term for any variety associated with a particular configuration of situational characteristics and purposes. Biber et al. (1999) in their *Longman Grammar of Spoken and Written English* distinguish between four major registers according to several situational characteristics (Table 1). Academic writing is characterized by a lack of interactiveness, online production, and shared immediate situation, its main communicative purpose being information, argumentation and explanation to a specialist audience through global dissemination.

Table 1. Major situational differences among four primary registers (Biber et al., 1999, p. 16)

	Conversation	**Fiction**	**News**	**Academic**
Mode	spoken	written (+ written dialogue)	written	written
Interactiveness and online production	yes	(restricted to fictional dialogue)	no	no
Shared immediate situation	yes	no	no	no
Main communicative purpose/content	personal communication	pleasure reading	information/ evaluation	information/ argumentation/ explanation
Audience	individual	wide-public	wide-public	specialist
Dialect domain	local	global	regional/ national	global

Biber and Conrad (2009, Chapter 5) present a more detailed description of academic prose, its situational characteristics and linguistic features. According to Biber and Conrad (2009, p. 32), more specialized written academic registers (sub-registers, also known as text types or genres) may differ along various parameters, e.g. intended audience, communicative purpose, and academic discipline.

Academic writing is among the most difficult registers for language users to master. Given the high cognitive demand placed on participants and the fact that exposure and use are generally limited to higher levels of education, it is worth emphasizing that even many native speakers (NSs) never achieve mastery in academic writing; for native and non-native users alike, then, academic writing skills represent the most advanced levels of writing proficiency. Even though secondary school educated students may have relatively little awareness of academic writing conventions, those who hope to succeed in tertiary education and beyond will need to learn to effectively communicate using this register.

Though a universal description of academic writing has not yet been agreed upon, even in a language as widely used as English, recent research in the field has aimed at describing advanced interlanguages by specifying the areas that are still problematic even for learners at the (very) advanced stages of language learning. In a recent overview of the field, Granger (2008) 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" (2008, p. 269). According to her, typical features are an 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" (2008, p. 269).

Advanced learners have also been found to struggle with the acquisition of linguistic phenomena that are optional and highly specific in the foreign/second language (L2) (DeKeyser, 2005, pp. 7 ff.), which are often located at the interfaces of linguistic subfields, e.g. at the syntax-semantics or the syntax-pragmatics interface (see e.g. Callies, 2009). As for academic writing, many of the observed difficulties seem to be 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, p. 112). In sum, many studies based on learner corpora have provided evidence that advanced learners of various L1 backgrounds have similar problems and face similar challenges on their way to near-native proficiency. In view of these similarities, the interlanguage of these learners can be conceived of as Advanced Learner Varieties (ALVs) (see Callies, 2013).

Yet, despite this growing interest in advanced proficiency, the field is still struggling with a definition and clarification of the concept of "advancedness". Recently, Ortega and Byrnes (2008) have discussed four partially overlapping global measures commonly used to operationalize advancedness: institutional status, standardized tests, late-acquired linguistic features, and a concept they call "sophisticated language use in context". It is this last measure that the authors particularly favour in which advancedness is conceptualized not only in terms of "purely linguistic accomplishments", but also — among other things — in terms of literacy, "choice among registers" and "voice" (2008, p. 8). This approach is in line with Shohamy's (2006) suggestion to discuss advanced proficiency in relation to "multiplicity", i.e. "multiple ways to 'know' and a variety of ways of being an 'advanced' language user" that are "embedded in diverse contexts and goals" (2006, p. 194); in this view, "language performance is judged not by abstract "native speaker" criteria but by various content- and context-related criteria" (2006, p. 194).

3. Validity of common SLA research assessment methods

Assessing the writing proficiencies of individuals or groups of language learners is clearly useful in various contexts, including SLA research where, for example, the aim is often to generalize patterns and distinctions between differing levels of language proficiency. The meaningfulness of any findings in this type of research, however, is dependent on the validity and the reliability of the measurement(s) used to inform the study (cf. Bachman, 1990; Fulcher & Davidson, 2007). In the pursuit of meaningful SLA research findings, then, care should be taken that a

"test score reflects an underlying ability accurately" and a "test's results are accurate, consistent and dependable" (Barker, 2010, p. 633).

In his pioneering argument towards a "unitary" view, Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (1989, p. 5; emphasis in original). This view of test validity rejects the notion of distinct types of validity, such as content or criterion, instead proposing that several appropriate sub-concepts, with *construct validity* considered to be primary, be taken into account when building a validity argument for a particular test use.

Though in theory this means that researchers should always provide details and validation arguments for the measures used to group or compare learners in their studies, Thomas (1994) observed in her review of L2 proficiency assessment methods in SLA research that information on the effects of varying levels of proficiency has in practice often been lacking. This inadequacy leads to clear limitations in the strength of these findings and, particularly, in the comparability of findings across studies. Thomas's analysis revealed four major groups of measurement techniques: institutional status, impressionistic judgment, standardized tests, and various forms of in-house assessment. Two of these, institutional status and standardized tests, overlap with Ortega and Byrnes' (2008) four measures of advancedness and, though late-acquired linguistic features and sophisticated language use are not specifically mentioned by Thomas (1994), in practice these have been dealt with by researchers to varying degrees in her remaining two categories.

Of the four types of L2 proficiency assessment methods Thomas identified, institutional status, i.e. identifying proficiency on the grounds of learners' position within some kind of "hierarchically-organized social structure" (1994, p. 317) is the least theoretically grounded. Objective and convenient as it is, this technique is limited in that it can only be expected to show mean intergroup variation. It often disregards the possibility, indeed the probability, of *intra*group variation in language proficiency. Within any given institutional level there are bound to be less proficient students who are more typical in many ways of the previous level, and more advanced students who are more characteristic of the next level; sole use of institutional status to define proficiency levels ignores these within-group differences. Though it may be a relatively quick and easy grouping method, it has relatively low validity with respect to its power to identify the more and less proficient learners, as there is a clear mismatch between institutional status and the true construct of interest: language proficiency (see also below for further discussion vis-à-vis corpus approaches).

Defining proficiency according to impressionistic judgment, e.g. on the basis of recommendation or "word of mouth" recruiting without further testing (e.g.

Coppieters, 1987) is also problematic due to the subjectivity of the judgments. In these cases, proficiency levels are determined by the professional experience and personal perceptions of raters, i.e. on their intuition. Though this type of assessment, which at least attempts to address the specific construct of language proficiency, may be considered to be more valid than institutional status, classification judgments tend to be less reliable in cases where the raters differ in experience or interpretation of the underlying construct and its relevant levels. Though reliability in rater judgment can generally be improved through rater training (cf. Fulcher & Davidson, 2007; Hughes, 2003), a clear definition of the relevant construct and the defining characteristics of its distinct levels are essential.

Standardized tests such as the *Test of English as a Foreign Language* (TOEFL), the *International English Language Testing System* (IELTS), and the *Pearson Test of English* (PTE) often report clear construct definitions and validity arguments that include relatively high reliability coefficients. It should be remembered, however, that validity is specific to the context in which the results will be interpreted: the relevance of the test to the goals of the research at hand must be carefully considered. Studies concerned with defining, describing, and understanding the characteristics of the highest levels of academic writing proficiency, for example, may consider the tightly constrained nature of the output elicited by typical task(s) in these tests to be inauthentic and, therefore, problematic. Undoubtedly, without a clear and common construct it is impossible to make meaningful generalizations based on the findings of a study. Therefore, when a test does not already exist that reasonably addresses the needs of a particular study, SLA researchers often find it necessary to design their own customized "in-house" tests; this is particularly the case when researchers are interested in specific linguistic features that may not be clearly addressed in existing tests.

In her 2006 follow-up research synthesis, Thomas noted a tendency of SLA researchers to disregard the importance of proficiency assessment in that measurement of L2 proficiency was considered "unnecessary, unreliable, or unrevealing" (2006, p. 294). At the same time, however, she indicates that improved evaluation methods and more specific study designs now lead to more valid results (as manifest in an increased variety and diversity of forms of assessment, especially in tests used for in-house assessment). Despite its wide use and necessity in cases where existing measures do not accurately reflect the construct of interest, Thomas's findings seem to indicate that in-house test development requires careful planning and consideration and, furthermore, that better assessment practices are increasingly being prioritized by SLA researchers. In order to be maximally effective, however, the measures resulting from these practices should not only be reliable and valid indicators of the construct researchers are most interested in, in-house tests should be theoretically grounded in a way that allows for meaningful comparisons

across studies. Despite this possible improvement in assessment practices among SLA researchers in general, without a clear and common view of the construct, SLA studies, including those of (advanced) academic writing, will continue to lack cross-study comparability.

4. Using the CEFR in academic writing research

In the search for a theoretical framework that will easily allow these cross-study comparisons, the recent popularity of the CEFR as a common ruler in SLA research makes it an important candidate. The CEFR takes a functional approach to language proficiency, describing each of the levels according to sets of "can-do" statements. The problematic nature of these can-do statements becomes obvious, however, when they are applied to a particular language as each language presents its users with a variety of lexico-grammatical means to express a particular communicative function. It turns out, then, that those statements are often too global and underspecified, and thus, of limited practical use for language assessors. Hence, there is a growing awareness among researchers of the need to specify the CEFR framework by developing more explicit descriptors anchored in language use (e.g. Hawkins & Buttery, 2010; Hawkins & Filipović, 2012).

The CEFR is also lacking with respect to the register of academic writing (Neff et al., 2008; Neff van Aertselaer & Bunce, 2011); in its current state, the CEFR is poorly equipped to benchmark written language samples and proficiency according to its six levels (A1 through C2). Though a large number of proficiency descriptors were empirically calibrated to the CEFR levels (Council of Europe, 2001; North, 2000), these validated descriptors mainly pertain to spoken language. Rather than ignore other aspects of language proficiency that had not yet been empirically validated, the CEFR was published in 2001 with so-called "illustrative descriptors" for other skills, including written production. The current CEFR descriptors for writing proficiency have not been empirically validated, nor do they claim to represent all aspects of written production that may be relevant in defining proficiency. In fact, they have largely been reconstructed from scales describing different skills entirely (Council of Europe, 2001, p. 61).

Despite its flaws, the CEFR has gained traction across Europe and, indeed, around the world as the fashionable way to compare and contrast actual and target language proficiency across users, courses, examinations, institutions, etc. It is, therefore, increasingly being used as a starting point in more specific SLA studies such as those concerning academic writing. Several papers in this issue report on recent studies that attempt to address the subjects of academic writing assessment and development using a common proficiency framework. For the most part,

these attempts fall into two categories of research: those relating specific language examinations to the CEFR, and those utilizing corpus-linguistic tools and methods to assess and better understand academic writing development.

4.1 Linking specific examinations to the CEFR

Following the CEFR itself, the Council of Europe released a manual "to help the providers of examinations to develop, apply and report transparent, practical procedures in a cumulative process of continuing improvement in order to situate their examination(s) in relation to [CEFR]" (2009, p. 1). Those who choose to adhere to the procedures outlined in the manual will follow five major stages: familiarization, specification, standardization training and benchmarking, standard setting, and validation. Each of these stages will be addressed in turn.

Prior to beginning the linking project, and particularly before starting the specification and standardization phases, all members of the linking committee should complete thorough familiarization procedures "to ensure that participants in the linking process have a detailed knowledge of the CEFR, its levels and illustrative descriptors" (2009, p. 10). Aside from simply reading the relevant sections of the framework itself, participants should complete familiarization techniques such as rating illustrative samples that have been benchmarked to various CEFR levels and qualitative discussion of relevant language samples and ratings with other committee members.

Once the committee members have undergone sufficient familiarization training, the specification stage of the linking process consists of a content analysis of the examination with respect to the CEFR in essence allowing the committee to "[profile] their examination in relation to CEFR categories and levels" (2009, p. 26). More specifically, this stage should include a clear and detailed description of the test (cf. Council of Europe, 2009, pp. 27–28, pp. 126–131) and a thorough content analysis based on this description, using various specification tools exemplified in the manual (pp. 28–34, pp. 131–152).

The standardization training stage is a natural extension of the familiarization stage, during which time the committee members consult exemplar performances and relate additional texts to the CEFR levels, generally in a group setting. Committee members discuss a number of performances and how they relate to the CEFR, and it is especially important at this time to ensure that all members achieve consensus on how to interpret the CEFR levels in relation to the relevant performance samples.

Different procedures are necessary for direct testing methods, where the CEFR level is judged holistically, than are suggested for indirect testing methods, which produce a numerical score relating to analytical scoring methods (in the case of

productive skills such as writing). While the benchmarking completed in the standardization stage may be sufficient to ensure that raters are working consistently in the case of direct tests, some method of standard setting will be necessary for indirect tests in order to determine appropriate cut-off scores relating to each relevant CEFR level (see Council of Europe, 2009, p. 57–87 for further information on standard setting).

Validation, the final stage in the linking process, "concerns the body of evidence put forward to convince the test users that the whole process and its outcomes are trustworthy" (2009, p. 90). As previously argued, all tests should be validated for a particular use. When linking a test to the CEFR, the validation argument should not be limited to the test at hand and its primary use; it should also provide relevant information allowing test users to make informed decisions on the usefulness of the link provided between the test and the CEFR.

Several papers in this issue present case studies of tertiary education settings in which various stages of this CEFR linking process have been carried out. Haines, Schmidt, Jansma and Lowie (this volume) underscore the importance of familiarization and specification in their report on their Dutch university's attempt to embed the CEFR in writing assessment procedures. Heaney (this volume) reports on the standardization process of a national school-leaving exam for bachelor students in Austria. Haapanen, McAnsh, Braidwood and Hollingsworth (this volume) describe the processes undertaken by a consortium of university language centers in Finland in order to increase transparency in English assessment using the CEFR. Zheng and Mohammadi (this volume) explore the construct validity of the writing tasks in a standardized test.

4.2 Corpus-linguistic analyses of CEFR levels

Corpora and corpus linguistic tools and methods have been used for some time in the study of L2 learning. Learner corpora are systematic collections of authentic, continuous and contextualized language use by L2 learners and are increasingly being used for language testing and assessment (LTA; e.g. Barker, 2010; Hawkins & Buttery, 2010; Hawkins & Filipović, 2012; Taylor & Barker, 2008). Corpora in general have the potential to increase transparency, consistency and comparability in the assessment of L2 proficiency. In recent overviews, Barker (2010) and Callies, Diez-Bedmar and Zaytseva (2012) discuss possible practical applications of NS reference and learner corpora in LTA. These can range from corpus-informed to corpus-based and corpus-driven uses, depending on the way corpus data are actually put into practice, the aims and outcomes for LTA, and the degree of involvement of the researcher in terms of data retrieval, analysis and interpretation. In *corpus-informed* applications, corpus data are either used as a reference source

of native-like, expert language usage, to provide test evidence, e.g. to inform test content, or to validate human rating. Language usage as included in NS reference corpora can thus be conceived of as an "abstracted corpus norm, which is an operationalisation of the native speaker norm" (Mukherjee, 2005, p. 16). For example, the *Pearson International Corpus of Academic English* (PICAE; Ackermann, De Jong, Kilgarriff, & Tugwell, 2011) was compiled as part of the development programme for the *Pearson Test of English Academic* to produce "a frequency list of the most common and pedagogically relevant collocations in written academic English discourse" which "can be used, for example, in lexicography, test item writing, and EAP [English for Academic Purposes] material development" (Ackermann, Biber, & Gray, 2011).

Another strand of research combines both the use of the CEFR to describe students' proficiency levels and Computer-aided Error Analysis (CEA; Dagneaux, Denness, & Granger, 1998) to analyse errors made by students at various CEFR levels, and to validate human rating (e.g. Diez-Bedmar, 2011; Thewissen, 2012).

In *corpus-based* approaches, learner language is explored and often compared to that of NSs to provide empirical evidence and either confirm or refute the hypothesis of a researcher. In contrast to corpus-informed approaches, corpus-based studies can yield information on the use of learner language in situations comparable to those of NSs which are not constrained by situational characteristics of task setting and test situation. The recent work by Hawkins and colleagues (2010; 2012) on criterial features in L2 English exemplifies the corpus-based approach. Their work aims at identifying linguistic descriptors to make information on CEFR proficiency levels more explicit by adding "grammatical and lexical details of English to CEFR's functional characterisation of the different levels" (Hawkins & Filipović, 2012, p. 5). The procedure involves comparing particular linguistic features as used by learners and NSs in two kinds of corpora: the *Cambridge Learner Corpus* (CLC; composed of exam scripts produced at different proficiency levels) and a corpus of NS English, the *British National Corpus* (BNC). Depending on similarities and differences in usage patterns across corpora, linguistic features acquire the status of either positive or negative linguistic properties respectively and are interpreted as criterial features that are "characteristic and indicative of L2 proficiency at each level, on the basis of which examiners make their practical assessments" (Hawkins & Filipović, 2012, p. 6).

Finally, *corpus-driven* approaches (in the sense of "data-driven") presuppose the least degree of involvement on the part of the researcher in that they rely on computer techniques for data extraction and evaluation. The role of the researcher in that case is to formulate questions and to draw conclusions derived from what corpus data reveal when subjected to statistical analysis. For example, Wulff and Gries (2011) propose to measure learners' accuracy in the use of

lexico-grammatical association patterns by means of corpus-driven methods, i.e. collostructional analysis. While such methods have not yet been widely used in LTA, this kind of approach is considered particularly useful for a text-centered, data-driven classification of proficiency levels based on linguistic descriptors typical of academic prose (see Callies & Zaytseva, this volume).

Notwithstanding obvious benefits, the use of corpora in LTA can be problematic and might even lead to misleading generalizations in case a number of criteria have not been satisfied. Generally speaking, a corpus is useful for LTA to the extent that the information it contains is reliable in terms of the language variety it represents and that it is useful with respect to test purposes. Thus, whereas both the CLC mentioned above and the most widely used learner corpus to date, the *International Corpus of Learner English* (ICLE; Granger, Dagneaux, Meunier, & Paquot, 2009) both represent learner language, they provide two different kinds of data. While the CLC is a collection of learner exam scripts across different proficiency levels, the ICLE samples the writing of advanced learners of English in the form of literary and argumentative essays. In the case of the CLC, all kinds of generalizations are bound to be derived from information contained in exam scripts and thus are tightly constrained by the task setting and possibly influenced by the test environment. Meanwhile, in order to enable corpus-based enquiries relevant for LTA, one has to make sure that the situation of language use created by the procedure of test taking actually suggests a possibility of using a language in a way that could be potentially revealing for researchers.

Furthermore, in the case of using the CLC for developing criterial features, the question arises as to whether the issue of a possible influence of variables other than L1 has been part of the discussion. Another related question concerns the extent to which the kind of comparative analyses suggested by Hawkins and colleagues are appropriate in terms of corpus comparability: language produced in an exam situation as a response to a particular task is compared with highly contextualized NS usage as represented in the BNC, which is different from the learner corpus in terms of the whole situation of language use and thus hardly makes it possible to compare task setting variables. Differences between texts produced by L1 and L2 writers may in fact turn out to result from differences in task-setting (e.g. prompt, timing, access to reference works, see Ädel, 2008), and possibly task-instruction and imagined audience (see Ädel, 2006, p. 201ff. for a discussion of corpus comparability).

Similar questions apply to the use of the ICLE. ICLE-informed generalizations in the context of LTA are valid provided that such variables as language proficiency, register/genre, task setting and imagined audience etc. have sufficiently been controlled and documented. However, L2 proficiency in particular has often been a fuzzy variable in learner corpus compilation and analysis. Due to practical

constraints it is often operationalized by "learner-centered" methods (Carlsen, 2012, p. 165) such as institutional status. In the compilation of the ICLE, learners' proficiency level was assessed globally by means of external criteria: learners were considered advanced because of their institutional status as "university undergraduates in English (usually in their third or fourth year)" (Granger et al., 2009, p. 11). However, the results of human rating of twenty essays per ICLE-subcorpus according to the CEFR levels (Granger et al., 2009, p. 12) suggest that the proficiency level of learners represented in the ICLE actually varies between (higher) intermediate to advanced. While some ICLE-subcorpora predominantly seem to include learners from either the CEFR's B2 (e.g. learners whose L1 is Chinese, Japanese, Tswana, and Turkish) or C1 proficiency levels (e.g. Bulgarian, Russian, and Swedish EFL learners), others show a higher degree of intragroup variability (e.g. Czech, German, and Norwegian students). Such individual differences often go unnoticed or tend to be disregarded in learner corpus analysis and are thus not reported in favour of (possibly skewed) average frequency counts. Recent studies confirm that global proficiency measures based on external criteria alone are not reliable indicators of proficiency for corpus compilation (Mukherjee, 2009; Callies, 2013). Such challenges tackled, language corpora present LTA experts with diverse possibilities for exploring learner language and L2 proficiency in general, and academic writing proficiency, in particular.

Papers presented in the second part of the issue, thus use corpus-linguistic methodology in the consideration and further specification of advancedness in academic writing. Using a longitudinal collection of texts produced by Dutch university students of English, alongside a NS control corpus, De Haan and Van der Haagen (this volume) explore the development of sophisticated language use. Verheijen, Los and De Haan (this volume) similarly investigate the role of information structure in advanced writing. Present-Thomas, Weltens and De Jong (this volume) discuss learner- and text-based methods of CEFR level classification, using basic corpus-linguistic methodology to explore each method. Finally, Callies and Zaytseva (this volume) introduce a new corpus of academic learner writing and describe its potential use in a text-centered, corpus-driven approach to assess advanced writing proficiency.

References

Ackermann, K., Biber, D., & Gray, B. (2011). An academic collocation list. Paper presented at *Corpus Linguistics 2011*, 20–22 July 2011, Birmingham, UK.

Ackermann, K., De Jong, J.H.A.L., Kilgarriff, A., & Tugwell, D. (2011). The Pearson International Corpus of Academic English (PICAE). Proceedings of *Corpus Linguistics 2011*, Birmingham.

Available from http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-47.pdf

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John 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* (pp. 35–53). Amsterdam: Rodopi.

Alderson, J.C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 633–645). New York: Routledge.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Callies, M. (2009). *Information highlighting in advanced learner English*. 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. (this volume). The *Corpus of Academic Learner English* (CALE) — A new resource for the assessment of writing proficiency in the academic register.

Callies, M., Diez-Bedmar, M.B., & Zaytseva, E. (2012). Using learner corpora for testing and assessing L2 proficiency. Manuscript submitted for publication.

Carlsen, C. (2012). Proficiency level — A fuzzy variable in computer learner corpora. *Applied Linguistics*, *33*(2), 161–183.

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language*, *63*, 545–557.

Council of Europe (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment (CEFR): A manual*. Strasbourg: Language Policy Division.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*, 163–174.

De Haan, P., & Van der Haagen, M. (this volume). Assessing the use of sophisticated EFL writing: A longitudinal study.

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

Department of Translation Studies, University of Tampere, Finland. (n.d.). *What is 'academic writing'?* Available from http://www.uta.fi/FAST/FIN/RESEARCH/acadwrit.html (accessed 12 September 2012).

Díez-Bedmar, M.B. (2011). Spanish pre-university students' use of English: CEA results from the University Entrance Exam. *International Journal of English Studies*, *11*(2), 141–158.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *1*(4), 253–266.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Granger, S. (2008). Learner corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics. An international handbook*, vol. 1 (pp. 259–275). Berlin and New York: Mouton de Gruyter.

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.

Haapanen, P., McAnsh, S., Braidwood, E., & Hollingsworth, R. (this volume). UniTIE: Towards a transparent description of learning outcomes for academic writing.

Haines, K., Schmidt, N., Jansma, P., & Lowie, W. (this volume). EMBEDding the CEFR in academic writing assessment: A case study in training and standardization.

Hawkins, J.A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*(1), 1–23.

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

Heaney, H. (this volume). Fleshing out CEFR descriptors at C1 and above for the assessment of academic writing in departments of English at Austrian universities.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Hulstijn, J.H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal*, *91*(4), 663–667.

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

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Mukherjee, J. (2005). The native speaker is alive and kicking — Linguistic and language-pedagogical perspectives. *Anglistik*, *16*(2), 7–23.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 203–230). Amsterdam: John Benjamins.

Neff, J., Bunce, C., Dafouz, E., Gallego, J., Rica, J.P., Genís, M., & McCabe, A. (2008). The question of descriptors for academic writing in the European language framework: A critical view. *Kalbotyra*, *59*(3), 213–221.

Neff van Aertselaer, J., & Bunce, C. (2011). The use of small corpora for tracing the development of academic literacies. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora* (pp. 63–84). Amsterdam: John Benjamins.

North, B. (2000). *The development of a common framework scale of language proficiency*. Theoretical studies in second language acquisiton, vol. 8. Brussels: Peter Lang.

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/Taylor & Francis.

Present-Thomas, R., Weltens, B., & De Jong, J.H.A.L. (this volume). Defining proficiency: A comparative analysis of CEF level classification methods in a written learner corpus.

Shohamy, E. (2006). Rethinking assessment for advanced language proficiency. In H. Byrnes, H. Weger-Guntharp, & K.A. Sprang (Eds.), *Educating for advanced foreign language capacities: Constructs, curriculum, instruction, assessment* (pp. 188–208). Georgetown: Georgetown University Press.

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohamy, & N.H. Hornberger (Eds.), *Encyclopedia of language and education*, *2nd edition, volume 7: Language testing and assessment* (pp. 241–254). New York: Springer.

Thaiss, C.J., & Zawacki, T.M. (2006). *Engaged writers, dynamic disciplines: Research on the academic writing life*. Portsmouth, NH: Boynton/Cook.

Thewissen, J. (2012). *Accuracy across proficiency levels and L1 backgrounds: Insights from an error-tagged EFL learner corpus*. Unpublished PhD dissertation, Université catholique de Louvain, Belgium.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*(2), 307–336.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J.M. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). Amsterdam: John Benjamins.

Verheijen, L., Los, B., & De Haan, P. (this volume). Information structure: The final hurdle? — The development of syntactic structures in (very) advanced Dutch EFL writing.

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

Zheng, Y., & Mohammadi, S. (this volume). An investigation into the writing construct(s) measured in Pearson Test of English Academic.